TESIS DOCTORAL

Métodos de valoración urbana

Para aspirar al grado de Doctor por la Universidad de Córdoba de

José Carlos Casas del Rosal

Dirigida por:

Dr. José María Caridad y Ocerin

Dra. Julia M. Núñez Tabales

Programa de doctorado de Ciencias Sociales y Jurídicas

Departamento de Estadística, Econometría, Investigación Operativa, Organización de Empresas y Economía Aplicada

Córdoba, 2017

TITULO: *MÉTODOS DE VALORACIÓN URBANA*

AUTOR: *José Carlos Casas del Rosal*

© Edita: UCOPress. 2017
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TÍTULO DE LA TESIS: Métodos de Valoración Urbana

DOCTORANDO/A: José Carlos Casas del Rosal

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

En un primer momento, se efectúa una revisión literaria de los diferentes métodos de valoración de inmuebles. No obstante, el núcleo principal de este trabajo proporciona un software para valorar viviendas, locales comerciales o naves industriales en cualquier punto de la geografía nacional, teniendo presente variables relativas a las características que definen el inmueble y al entorno en el que se encuentran ubicados. Dicho software es aplicado en el apartado empírico a diferentes tipologías de inmuebles de la ciudad de Sevilla con excelentes resultados.

Como paso previo a la captura de los datos del análisis, se efectuó un estudio minucioso de los principales portales inmobiliarios que operan a nivel nacional, teniendo presente aspectos como número de inmuebles ofertados, tráfico web o características informadas.

La temática seleccionada ha dado lugar a varias contribuciones en congresos científicos –suscitando enorme interés entre los presentes-, así como diversas publicaciones en revistas científicas indexadas.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, ____ de ____JULIO_____ de ____2017_____

Firma del/de los director/es

Fdo.:_____ Fdo.: _____

A Carmen y David, sin los que este trabajo no hubiera sido posible.

Y al resto de mi familia, en especial al gran Don Antonio,
que nos dejó durante la redacción de esta tesis.

TQT

Agradecimientos

A los Doctores D. José María Caridad y Ocerin y Dña. Julia M. Núñez Tabales, directores de la presente tesis, por su continuo apoyo y sus determinantes contribuciones.

A mis amigos y compañeros del área de Estadística e Investigación Operativa, por su enorme apoyo y sus valiosos consejos, en especial a Rafi, Pepe y Manolo, a quienes siempre encuentro cuando lo necesito.

A Idealista.com, por el desinteresado apoyo dado para la elaboración de esta tesis. Su aportación ha sido fundamental para la viabilidad de este trabajo.

A los doctores Orlando Arencibia y Petr Sed'a y a los compañeros de la facultad de ciencias económicas de Ostrava, gracias por haberme acogido y hacerme sentir como en mi propio hogar.

A las doctoras Eva Rublikova y Maria Cristina Canavarro Teixeira, por la revisión desinteresada de este trabajo.

A Carmen, mi esposa y mejor amiga, por su infinita comprensión.

A mis padres, hermanos y demás familia, que nunca dejan de apoyarme.

Índice

ÍNDICE DE CONTENIDO

Índice

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1.Introducción

La asignación del precio de un bien, cuando éste desea ser posicionado en el mercado con éxito, es una labor que, en ocasiones, es de una elevada dificultad, fundamentalmente si el bien y el mercado se caracterizan por una fuerte heterogeneidad.

Un claro ejemplo de un mercado en el que esta tarea es especialmente difícil, es el mercado inmobiliario, y aún más, un mercado inmobiliario, como el español, que ha sufrido en los últimos años grandes cambios.

Agentes de la propiedad inmobiliaria, tasadores, constructores, entidades financieras, aseguradoras o incluso propietarios particulares que deseen vender o alquilar una propiedad, se encuentran con la difícil tarea de asignar un valor que reúna las condiciones de representatividad y adecuación.

Este precio debe ser coherente con las características que definen el inmueble como son lugar geográfico en el que se encuentra, el tamaño y otras características de las que dispone como cochera o terraza. Es definitiva, debe ser representativo del inmueble que se está valorando.

Por otro lado, si el valor del inmueble difiere significativamente de las valoraciones de los inmuebles de su zona, éste puede situarse en el mercado en una posición en la que sea prácticamente imposible su venta o alquiler, o, por el contrario, transferirse a un precio

por debajo del valor por el que hubiera podido venderse, generando un elevado coste de oportunidad. En definitiva, el precio debe estar adecuado a los precios de los inmuebles de su mercado local.

El objetivo principal de este trabajo es el de proporcionar una herramienta que facilite una correcta decisión en la valoración de los inmuebles, en cualquier punto de la geografía nacional, teniendo en cuenta los factores relativos a las características que los definen, y el mercado inmobiliario de su entorno.

Para alcanzar este objetivo era necesario, en primer lugar, habilitar un mecanismo de recogida de la información de los inmuebles a la venta o en alquiler en España. El hecho de imponerse, como meta, un marco geográfico tan extenso, impedía recurrir a agencias de la propiedad inmobiliaria a nivel local.

El problema añadido de un mercado inmobiliario tan cambiante, debido a la especial situación en la que estábamos inmersos desde 2008, hacía necesario, además, disponer de una fuente de información dinámica, ya que la valoración de un inmueble podía sufrir grandes cambios en breves períodos de tiempo.

Eso nos llevó a analizar, en profundidad, el mercado de los portales web inmobiliarios, que se desarrolla en el capítulo 5 de este trabajo. Tras este minucioso análisis, llegamos a la conclusión, por oferta, tráfico web y representatividad, que el portal www.idealista.com era el que podía disponer de una mayor fuente de información, tanto en número de inmuebles como en características de las que se informaba, que, además, incluía tanto a particulares como a agencias de la propiedad inmobiliaria.

El siguiente paso consistía en recoger la información de una forma eficiente y rápida, lo que impedía una recogida manual a través de la navegación en su portal, que contenía millones de inmuebles. Así, nos pusimos en contacto con la empresa propietaria del portal, Idealista, S.A., que disponía de un sistema de recogida de datos, implementado para ofrecer un servicio a sus clientes profesionales, agentes de la propiedad inmobiliaria.

De forma desinteresada, nos proporcionaron acceso a la información contenida en su servidor a través de una clave de acceso, a través de la cual podríamos realizar peticiones de información desde su servidor; sin la que esto hubiera sido imposible. Este acceso

disponía de limitaciones que pretendían, por un lado, evitar el colapso de sus sistemas por un elevado número de peticiones, y por otro, proteger su valiosa información. Restricciones que entendimos y asumimos, agradeciendo su enorme contribución a este trabajo.

El siguiente paso era implementar un software que permitiera recoger de forma automática la información necesaria del servidor de la empresa, desarrollando, además, un algoritmo de búsqueda que verificase las restricciones impuestas. Este software, además, debía transformar la información recibida en una base de datos que permitiera su procesamiento.

Por otro lado, era necesario proveer a este programa de la capacidad de añadir actualizaciones a las búsquedas realizadas, que permitieran estudiar la evolución de la oferta de inmuebles y el cambio en sus precios.

Una vez alcanzado este objetivo, disponíamos de una herramienta capaz de recoger gran cantidad de información relativa a la oferta de inmuebles de una zona y transformarla en un formato exportable.

Llegados a este punto, decidimos dotar al programa creado, de la capacidad de realizar todo tipo de estudios estadísticos de los datos, para lo que decidimos conectar el programa, realizado en lenguaje JAVA, con el potente motor de cálculo de R. Por lo que lo siguiente era decidir las técnicas a utilizar para el análisis de la información, y en particular, para la estimación del precio de un inmueble.

Para ello, en primer lugar, se analizaron minuciosamente todos los métodos de valoración inmobiliaria, tras lo que se consideró el método econométrico de comparación como aquél que podría dar los mejores resultados. Este método es comúnmente denominado metodología de precios hedónicos, y puede consultarse una extensa revisión de éste en el capítulo 2 de este trabajo. Ésta, ha sido ampliamente utilizada por numerosos autores, desde mediados del siglo pasado, para estimar la valoración de bienes, y en particular, el precio de un bien heterogéneo, esto es, conformado por un conjunto de características que lo definen.

A continuación, y analizando también la literatura existente, se procedió a estudiar en profundidad una disciplina de la inteligencia artificial, las redes neuronales, y, en particular, las construidas para la estimación del valor de una variable en función de aquellas de las que depende. Una extensa revisión puede consultarse en el capítulo 3.

No obstante, la construcción de modelos para la estimación requiere de un tratamiento previo de la información, de modo que se pueda determinar la existencia de observaciones, denominadas influyentes, que, por sus características únicas y diferenciadoras, puedan desvirtuar las estimaciones realizadas, pero que, sin embargo, no son representativas de ningún comportamiento a estudiar.

En el mercado inmobiliario, una observación influyente se traduce en un inmueble que dispone de un valor significativamente distinto, en alguna de sus características, de los observados en el resto. Esto puede significar que el inmueble se encuentra mal posicionado en el mercado en cuanto a precio, o a una clara desproporción entre los valores de sus características. Estamos ante, inmuebles que interesa analizar en profundidad.

Esta reflexión nos llevó a estudiar en profundidad los métodos de detección de valores atípicos, y en particular, de observaciones influyentes asociadas a modelos de regresión. Su revisión, puede encontrarse en el capítulo 4.

En este mismo capítulo se propone un método de detección de observaciones influyentes que ha demostrado, en las simulaciones realizadas y después en los estudios de campo, ser de gran utilidad y efectividad. Este método se basa en analizar la variación que se produce en el criterio de información de un modelo de regresión, al eliminar del conjunto de datos, una observación influyente.

Por último, añadimos un amplio análisis descriptivo univariante y bivariante de los datos, y aprovechando la disponibilidad de las coordenadas geográficas de los inmuebles, también la posibilidad de realizar mapas de situación de éstos con información de alguna de las características que los definen. Estos mapas se construyen sobre imágenes satélite de la zona en la que se ha realizado el estudio.

Todo esto fue implementado en lenguaje R a través de las funciones ampliamente descritas en el capítulo 7, dando lugar al software desarrollado, que hemos denominado INMODATAANALIZADOR. Una amplia documentación de este programa ha sido incluida en el capítulo 6.

Por último, nos propusimos utilizar las funcionalidades implementadas por el software para realizar el estudio de la oferta inmobiliaria de una zona concreta. Optamos por una gran ciudad española que, además, es capital de su comunidad autónoma, Sevilla. Para ésta se realizó, por un lado, un estudio del mercado de viviendas a la venta de la ciudad, y por otro, un análisis de los locales comerciales y las naves industriales de la conocida como aglomeración urbana de Sevilla.

Debido a la escasa presencia de trabajos existente sobre valoración de locales comerciales, y lo inédito de construir un modelo de estimación de precios de naves industriales, ha sido necesario analizar con detenimiento qué variables, de entre las disponibles, influyen de manera significativa sobre el precio de este tipo de inmuebles. Todo esto puede consultarse en el capítulo 8.

Introduction

The allocation of the price of a good when it is wished this to be successfully positioned in the market is a high difficult work sometimes, mainly if the good and the market are characterized by a strong heterogeneity.

The real estate is a clear example of a market in which this task is especially difficult, even more the Spanish real estate market which has undergone great changes in recent years.

Real estate agents, appraisers, builders, financial institutions, insurers or even private owners who want to sell or rent a property find the difficult task of assigning a value that meets the conditions of representativeness and adequacy.

This price must be consistent with the characteristics that define the property such as geographical location, size and other characteristics as if it has as a garage or a terrace. Definetly, it must be representative of the property that is being valued.

On the other hand, if the value of the property differs significantly from the valuations of the properties in its area, it may be placed on the market in a position where it is practically impossible to sell or rent it, or, on the contrary, being transferred at price below

the value that could have been sold and generating a high opportunity cost. In short, the price should be adequate to the prices of the real estate of the local market.

The main objective of this work is to provide a tool that facilitates a correct decision in the valuation of the real estate in any point of the national geography taking into account the factors related to the characteristics that define it and the real estate market of its environment.

Firstly, it was necessary to enable a mechanism for collecting information on real estate for sale or for rent in Spain in order to achieve this objective. The imposition, as a goal, of such a large geographic framework prevented the use of real estate agencies information at the local level.

The added problem of such a changing real estate market due to the special situation in which we were immersed since 2008 also required a dynamic source of information as well since the valuation of a property could undergo major changes in brief periods of time.

This led us to analyze the market for real estate web portals in depth, what is developed in chapter 5 of this work. After this thorough analysis, we concluded by offer, web traffic and representation that the www.idealista.com portal was the one that could have a greater source of information, as much in number of properties as in characteristics informed, it also included both individuals and real estate agencies.

The next step was to collect the information in an efficient and fast way, which prevented a manual collection through surfing this portal, which contained millions of real estate. Thus, we contacted the portal's company owner, Idealista, S.A., which had a data collection system implemented to offer a service to its professional clients, real estate agents.

Disinterestedly, they provided us access to the information contained in their server by a password through which we could make requests for information from their server, otherwise it would have been impossible. This access had limitations that sought, on the one hand, to avoid the collapse of their systems by a high number of requests, and on the

other, to protect their valuable information. We understood and assumed their restrictions thanking them for their great contribution to this work.

The next step was to implement a software that let us automatically collect the necessary information from the company's server, also developing a search algorithm that verified the restrictions imposed. This software, moreover, had to transform the information received in a database that allowed its processing.

On the other hand, it was necessary to provide this program with the ability to add updates to the already made searches, which allowed to study the evolution of the real estate offer and the change in prices.

Once this objective was reached, we had a tool capable to collect a large amount of information regarding the real estate offer in an area and transform it into an exportable format.

At this point, we decided to provide the created program with the ability to perform all statistical studies of the data, so we decided to connect the program, made in JAVA language, with the powerful R calculation engine. The next thing was to decide the techniques to be used for the analysis of the information and, in special, for the estimation of the price of a property.

In order to do this, first, all the methods of real estate valuation were thoroughly analyzed, after which the econometric method of comparison was considered as the one that could give the best results. This method is commonly called hedonic price methodology and an extensive review can be found in Chapter 2 of this paper about it. This has been widely used by numerous authors since the middle of the last century to estimate the valuation of goods and, in particular, the price of a heterogeneous good, that is, made up of a set of characteristics that define it.

Then, and also analyzing the existing literature, we proceeded to study in depth a discipline of artificial intelligence, neural networks and, specially, those constructed for the estimation of the value of a variable in function of those on which it depends . An extensive review can be found in Chapter 3.

However, the construction of models for estimation requires a prior treatment of the information, so that it is possible to determine the existence of so-called influential observations, which, due to their unique and differentiating characteristics, may detract from the estimates made, but are not representative of any behavior to be studied.

In the real estate market, an influential observation translates into a property that has a significantly different value in some of its characteristics from those observed in the rest. This may mean that the property is poorly positioned in the market in terms of price, or a clear disproportion between the values of its characteristics. We are before real estate that interests to analyze in depth.

This reflection led us to study in depth the methods of detection of atypical values and influential observations associated with regression models particularly. Its review can be found in Chapter 4.

In this same chapter, a new method of detecting influential observations is proposed, which has been shown to be very useful and effective in the simulations carried out and in the field studies later. This method is based on analyzing the variation occurring in the information criterion of a regression model by eliminating an influential observation from the data set.

Finally, we add a broad univariate and bivariate descriptive analysis of the data and, taking advantage of the availability of the geographical coordinates of the properties, the possibility of mapping the situation of these with information of some of the characteristics that define them as well. These maps are constructed on satellite images of the area in which the study was performed.

All this was implemented in R language through the functions broadly described in chapter 7, giving rise to the software developed, which we have called IMMODATAANALIZADOR. Extensive documentation of this program has been included in Chapter 6.

Finally, we proposed to use the functionalities implemented by the software to perform the study of the real estate offer of a specific area. We chose a large Spanish city that, moreover, is capital of its autonomous community, Seville. For this, a study of the

housing market for sale in the city was carried out, and on the other hand an analysis of the commercial premises and industrial buildings of the well-known urban agglomeration of Seville, which can be consulted in Chapter 8.

2.Métodos de valoración. Metodología de precios hedónicos

2.1. Métodos de valoración de inmuebles

Un bien inmueble es una entidad física constituida por dos grandes partes: el suelo y la edificación. Cada una de estas partes se pueden caracterizar a través de un conjunto de atributos.

La valoración inmobiliaria se define como la valoración de activos cuyo soporte es un bien material e inmobiliario. Pueden distinguirse dos tipos: valoración rural y urbana. En este trabajo nos centraremos en esta última. En esta definición pueden también incluirse otros elementos como los derechos reales – usufructo, servidumbres… etc. – y las cargas derivadas de la transacción como son las hipotecas. (Guadalajara, 2014)

Por tanto, el valor de un inmueble es, aquel que el mercado establece de acuerdo a sus características constructivas y cualitativas, así como a las características del entorno en el que este se encuentra. De esta forma, es fundamental para determinar su valor, analizar, además, en qué medida sus características son idóneas y se adaptan a los requerimientos del mercado en el que se encuentra.

En un sentido más específico, podemos definir otros valores inmobiliarios:

- Valor de mercado: valor más probable al que se venderá un inmueble en un mercado con un equilibrio racional entre la oferta y la demanda.
- Valor inversión: Valor subjetivo que debe tener el inmueble para un inversor, teniendo en cuenta sus requerimientos de beneficio.
- Valor futuro: Aquel valor que incluye las expectativas de futuro incremento de su valor debido a condiciones externas al mismo.
- Valor en renta: Es valor que tiene en cuenta, además, la capacidad para generar rentas, como por ejemplo inmuebles que se encuentran en régimen.
- Valor de fondo de comercio: En el caso de que en el inmueble se desarrolle una actividad económica, es el valor que tiene en cuenta, además la rentabilidad económica que ésta genera.
- Valor intrínseco: El relativo al valor de coste de fabricación, en el que se incluye la construcción del edificio, la adquisición de suelo y la promoción, en su caso.
- Valor de afección: Valor que para el propietario tiene el inmueble según un criterio subjetivo, sin asumir criterios económicos.

Fundamentalmente, pueden distinguirse dos métodos de valoración, los métodos comparativos y los analíticos. El primero de los métodos se basa en comparar el inmueble a valorar con inmuebles con características similares. El segundo, trata de valorar el inmueble a través de criterios técnicos de construcción o de valoración del suelo.

Según Guadalajara (2014), las causas que pueden motivar la aplicación de uno u otro método de valoración son las siguientes:

- Disponibilidad de información.
- Naturaleza del bien a valorar.
- Finalidad de la valoración.
- Definición de valor utilizada.

Nos centraremos en este trabajo en los métodos comparativos. Una explicación detallada de los métodos analíticos puede consultarse en Núñez (2007).

2.2. Métodos de valoración por comparación

Las técnicas de valoración del precio de mercado de un inmueble tienen como objetivo fundamental estimar este valor por comparación con otros inmuebles con características similares a las del objeto de la valoración.

Para ello, además del precio de mercado de los inmuebles con los que es comparable, puede utilizarse otros valores como el precio de oferta o el precio de tasación. (Guadalajara, 2014)

Para la utilización de estos métodos es necesaria la disponibilidad de información del mercado inmobiliario del entorno en el que se desea realizar la valoración, lo que en ocasiones no es posible, ya que se necesitará para ello información relativa a un elevado número de ellos.

Según Guadalajara (2014) la herramienta fundamental para resolver con éxito una valoración inmobiliaria mediante el método de comparación es la ejecución de un correcto estudio del mercado de oferta del entorno. Las características fundamentales de este estudio deben ser:

- Disposición de una muestra de tamaño elevado, aunque esto suele traducirse en un alto coste. Esta necesidad ha sido resuelta en este trabajo como se analizará posteriormente.
- Elaboración de una base de cada uno de los inmuebles disponibles que contenga al menos: localización, tipología edificatoria, submercado, grado de conservación, superficie, altura, calefacción, distancia al centro urbano, características del entorno, presencia de esquina – en locales comerciales) -, distribución… etc.
- Verificación de la información descartando de ésta aquellos con características anómalas, que pueden desvirtuar la estimación.
- Estratificación de la muestra según el uso – viviendas, locales comerciales… etc. -, la tipología edificatoria, la localización y la calidad edificatoria.

Podemos distinguir, tres métodos de comparación: los métodos sintéticos, el método comparación de dos funciones de distribución (DFD) y el método econométrico o de regresión. Los dos primeros los comentaremos brevemente, pero puede encontrarse una descripción detallada en Guadalajara (2014).

Caballer (2008) clasifica los métodos sintéticos en métodos de comparación espacial, en los que se define una región alrededor del inmueble a valorar, y se analiza el precio de venta - o de mercado - actual de los inmuebles contenidos en dicha región con características similares; y el método de comparación temporal, en el que se analiza el precio que el bien a valorar ha tenido, en las distintas transacciones llevadas a cabo, a lo largo del tiempo.

Estos métodos requieren de gran cantidad de información de inmuebles a la venta y de un número elevado de transacciones previas en el de comparación temporal, lo que lo convierte en un método de difícil aplicación en muchas situaciones.

Por el contrario, el método de comparación de funciones de distribución, conocido como el método beta y desarrollado por Ballestero (1973) se basa en estimar el valor del

inmueble a partir del valor tomado por una característica fuertemente relacionada con éste, que será la variable explicativa del valor, como puede ser el ancho de la fachada de un local comercial o la distancia al centro de la ciudad.

Para su aplicación es necesario suponer que las funciones de densidad de ambas distribuciones se ajustan bien a alguna función conocida, y que ambas variables – variable valor y variable explicativa -. están fuertemente relacionadas. Las funciones de densidad más utilizadas son: beta, Normal, triangular, rectangular y trapezoidal.

Para ambas variables se determina su función de densidad, a partir de los valores disponibles, se confirma la fuerte asociación entre ambas y a continuación se calcula el valor de la variable valoración que verifique el valor de su función de distribución coincida con el de la variable explicativa, - o su complementario si la relación es inversa.

Las limitaciones fundamentales de este método son la suposición de que se puede estimar el valor de un inmueble con exactitud a partir de una característica concreta de éste, y, además, que, para construir correctamente la función de densidad de ambas variables, debemos disponer también de un elevado número de datos.

2.3. Fundamentos y antecedentes de la metodología de precios hedónicos

El tercero de los métodos por comparación es el método econométrico o de regresión, habitualmente conocido como el método de precios hedónicos.

Desde hace siglos, las personas han intentado determinar métodos que permitan estimar de forma objetiva el valor de un bien. Esta tarea no es sencilla cuando el bien a valorar es heterogéneo.

Un bien heterogéneo es aquél que se transmite indivisiblemente junto a un conjunto de características que determinan su valor. Lancaster (1966) determina que el consumidor de un bien ejerce sus preferencias sobre las características del mismo, y no en el bien en sí mismo.

Por tanto, si un bien dado está formado por un conjunto de características que determinan las preferencias de los consumidores, la valoración de éste puede realizarse a partir del análisis de las características que lo componen y la valoración que éstas tienen para los consumidores. Este es el punto de inicio de la metodología de precios hedónicos (MPH).

Desde principios del siglo XX hasta la actualidad se han sucedidos trabajos que tienen como objetivo principal valorar un bien heterogéneo a través de sus características.

Según el trabajo realizado por Colwell y Dilmore (1999), uno de los primeros trabajos que se utilizó la metodología de los precios hedónicos fue el de Haas (1922). Éste analizó el precio de venta por acre de un conjunto de 160 granjas de Minnesota, a partir de características como el estado de la carretera de acceso, la distancia al núcleo urbano, el tamaño del núcleo urbano más próximo o la productividad del suelo.

Posteriormente, en, Wallace (1926) se presenta un estudio similar con datos de 99 granjas del estado de Iowa[1].

En 1939 se publica *The dynamics of automobile demand*. (Court, 1939). Este trabajo se realizó tras el encargo de General Motors para explicar cómo la subida de los precios de un automóvil viene dada por el incremento en la calidad de los mismos.

En los años siguientes se sucedieron algunas publicaciones, sin apenas repercusión, hasta que los trabajos de Griliches (1961) y Rosen (1974), desarrollaron la metodología de los precios hedónicos, aportando los soportes necesarios para la aplicación de la teoría econométrica.

No obstante, antes, en 1967 destaca la aportación de Ridker y Henning, quienes, según Núñez (2007), aplicaron por primera vez la metodología de los precios hedónicos al

[1] Los datos recogidos por Wallace pueden consultarse en la librería *Agridat* del lenguaje R con el nombre *wallace.iowaland*, en ella pueden observarse datos como la longitud y la latitud, el porcentaje sembrado de maíz o de grano pequeño…etc.

mercado inmobiliario. En este estudio, los autores, demostraron empíricamente que la polución afecta negativamente al precio de un inmueble.

Fue a partir de 1974 cuando el desarrollo de la MPH experimentó un mayor auge, momento a partir del cual se publicaron, y siguen aún hoy publicándose, multitud de trabajos que han aplicado esta metodología a todo tipo de situaciones en las que se desea determinar el valor de un bien heterogéneo a partir de las características de éste.

Las aplicaciones al mercado inmobiliario en particular también han sido numerosas distinguiendo por un lado los estudios cuyo objetivo es determinar el precio de la vivienda, utilizando características estructurales o de localización y por otro lado los que tienen como objetivo obtener índices de precios.

Cabe destacar las aportaciones de Caridad y Brañas, (1996) en la aplicación de esta metodología en el mercado inmobiliario español.

En la Tabla 1 se muestran las principales aportaciones a la metodología de precios hedónicos, haciendo especial hincapié en las aportaciones de ésta en la estimación del precio de los inmuebles.

Tabla 1. Principales aportaciones a la MPH

Autor/es	Año	Ámbito y lugar de aplicación
Hass	1922	Estimación del precio de las granjas en Minnesota
Wallace	1926	Estimación del precio de las granjas en Iowa
Court	1939	Precio de los automóviles según calidad en Michigan
Griliches	1961	Desarrollo de la metodología
Ridker y Henning	1967	Influencia de la polución en el precio de la vivienda
Rosen	1974	Desarrollo de la metodología
Caridad y Brañas	1996	Mercado inmobiliario español

Fuente: Elaboración propia.

Por otro lado, las características utilizadas para explicar el precio de un inmueble son muy amplias. Pueden distinguirse dos grupos: las características estructurales como pueden ser las dimensiones, la antigüedad o determinadas calidades; y las características relacionadas con la localización del inmueble como puede ser distancia al centro de una ciudad, a parques y jardines o a determinados servicios públicos.

Se han utilizado numerosas características en los últimos años como regresoras de los precios de los inmuebles. Cabe destacar las siguientes aportaciones.

Núñez (2007) utiliza características para estimar el precio de un inmueble variables como la superficie, la antigüedad, la ubicación, gastos de comunidad o la existencia o no de trasteros o garajes, y halla significación en cada una de ellas para explicar éste.

Kong, Yin, y Nakagoshi (2007) estudiaron la influencia de la cercanía de los espacios verdes y las características de éstos sobre el precio de los inmuebles. También analizaron la cercanía a centros educativos.

Pope (2008) midió la influencia que la información que tiene el comprador sobre el ruido de un aeropuerto tiene sobre el precio de compra final del mismo, cuantificando el coste marginal que ésta tenía para el vendedor.

Gibbons y Machin (2008) analizaron la influencia en el precio de un inmueble de la red de transportes y el índice de criminalidad de una ciudad.

Bayer, Keohane, y Timmins (2009) cuantificaron la incidencia del grado de polución de una ciudad en el precio de compra de la vivienda.

Fuerst y McAllister (2011) observaron variaciones en el precio de un inmueble en función de que éste tuviera o no, la certificación energética, tanto en su precio de venta como de alquiler.

Wang, Potoglou, Orford, y Gong (2015) demostraron en una muestra muy amplia, cómo la existencia de paradas de autobús cercanas actúa positivamente en el precio de la vivienda, y el número de éstas tiene repercusión sobre el precio de ésta.

En la tabla siguiente se muestra un cuadro – resumen de las características utilizadas para la estimación del precio de un inmueble. Tabla 2.

Tabla 2. Algunas características usadas como variables regresoras del precio

Autor/es	Año	Características usadas
Núñez	2007	Superficie, antigüedad, gastos de comunidad, existencia de trastero o garaje
Kong, Yin y Nakagosi	2007	Cercanía de espacios verdes y de centros educativos
Pope	2008	Información que tiene el comprador sobre el ruido de un aeropuerto cercano
Gibbons y Machin	2009	Red de transportes e índice de criminalidad
Bayer, Keohane y Timmins	2009	Grado de polución de una ciudad
Fuerst y McAllister	2011	Estar en posesión o no, de la certificación energética
Wang, Potoglou, Orford, y Gong	2015	Existencia de paradas de autobús cercanas, así como el número de éstas

Fuente: Elaboración propia.

2.4. Metodología

A continuación, se desarrolla la metodología a seguir para la construcción de un modelo de regresión. Para ello se han tomado como referencia los libros de Caridad (1998), Gujarati y Porter (2011) y Montgomery, Peck, y Vining (2015).

2.4.1. Especificación del modelo

Sea Y una variable aleatoria continua que denominaremos variable dependiente o endógena, que es la variable que se desea analizar.

Sean también un conjunto de variables aleatorias $X_1, X_2, \dots, X_k$, que denominaremos independientes o exógenas, de las que asumiremos que ejercen determinada influencia sobre la variable dependiente Y.

Un modelo de regresión es una forma expresión de la forma:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

Donde f es una función real que a partir de valores dados para las variables independientes devuelve un valor, que es la estimación del valor dado por Y para dichos valores. Es decir:

$$\hat{Y} = f(X_1, X_2, \dots, X_k)$$

La variable ε se denomina error, residuo o conjunto de perturbaciones aleatorias, y es la variable que recoge el porcentaje de variabilidad de la variable Y no explicada por $f(X_1, X_2, \dots, X_k)$.

En el caso que nos ocupa, la variable dependiente Y representa un valor determinado de un bien heterogéneo, como puede ser su precio de venta. Este valor puede ser explicado por un conjunto de características intrínsecas a éste que influyen en su valor, y que son medidas a través del conjunto de variables independientes $X_1, X_2, \dots, X_k$.

En ocasiones, y con gran frecuencia en modelos hedónicos, es necesario incluir, en el modelo de regresión, variables explicativas que son de carácter cualitativo. Es evidente que por su naturaleza no pueden ser añadidas como una variable más.

Supongamos una variable cualitativa con p modalidades $m_1, m_2, \dots, m_p$ que se desea incluir como variable explicativa. Para ello generaremos p variables dicotómicas, también llamadas variables ficticias, indicadoras de la modalidad o *dummies*.

Estas variables se definen de la siguiente forma:

$$D_i = \begin{cases} 1 & si\ la\ varible\ toma\ la\ modalidad\ m_i \\ 0 & en\ caso\ contrario \end{cases} \quad con\ i = 1, \dots, p$$

De forma que cualquier individuo de la muestra tomará el valor 1 en una de las variables creadas y 0 en el resto.

Una vez creadas las variables *dummies*, éstas son incluidas en el modelo como una variable más teniendo en cuenta que a través del modelo podremos analizar la influencia que sobre la variable dependiente tiene que el individuo verifique esa modalidad concreta.

No obstante, la introducción en el modelo de todas las variables *dummies* creadas para una variable cualitativa, generaría una dependencia lineal exacta entre éstas, también

denominada multicolinealidad exacta, que impediría la estimación del modelo de regresión, como veremos más adelante. Por tanto, para una variable cualitativa con p modalidades, se generarán p variables *dummies*, de las que sólo se incluirán en el modelo $p-1$. La modalidad correspondiente a la variable *dummy* que queda fuera del modelo se denomina categoría base y puede ser elegida aleatoriamente.

Una vez analizadas las variables a incluir en el estudio, y tras recoger y analizar los datos, se elegirá la forma funcional f, que nos permita relacionar el conjunto de atributos o características con el valor del bien a estimar.

La elección de la forma funcional debe realizarse teniendo en cuenta la teoría subyacente al modelo que se desea construir. Se han realizado numerosos trabajos sobre la valoración de inmuebles, y en la mayoría de ellos, la forma funcional más utilizada es la lineal.

Es poco realista pensar, y así se afirma en Sanjuán, Hurlé, Pérez, y Royo, (2004) que el valor de un bien tenga una dependencia constante de sus características, de forma que el valor que aporte al bien una unidad más de una característica determinada sea independiente del número de unidades que se estén adquiriendo.

No obstante, si el rango de valores en el que se mueva la unidad no es amplio, esta hipótesis podría asumirse. No obstante, la solución a este problema más habitual, es la de hallar aproximaciones de los modelos no lineales mediante transformaciones logarítmicas de algunas de las variables en estudio. De esta forma, los modelos más habitualmente usados son los mostrados en la Tabla 3:

Tabla 3. Formulaciones linealizadas usuales.

Modelo	Expresión
Función lineal	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$
Función semilogarítmica	$ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$
Función doble logarítmica	$ln(Y) = \beta_0 + \beta_1 ln(X_1) + \beta_2 ln(X_2) + \cdots + \beta_k ln(X_k) + \varepsilon$
Función lineal logarítmica	$Y = \beta_0 + \beta_1 ln(X_1) + \beta_2 ln(X_2) + \cdots + \beta_k ln(X_k) + \varepsilon$

Fuente: Elaboración propia

Los test que nos permiten determinar la idoneidad de la forma funcional escogida requieren la verificación de hipótesis sobre el modelo que analizaremos en la fase de validación, por lo que retomaremos el tema en ese punto.

A partir de este momento, vamos a considerar el modelo función lineal, ya que todos los modelos expuestos son equivalentes en cuanto a la estructura de los mismos. Basta con realizar el cambio especificado en las variables del modelo.

2.4.2. Estimación del modelo

Una vez seleccionada la forma funcional, nuestro siguiente objetivo es estimar los coeficientes del modelo.

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$$

Construyamos, en primer lugar, el modelo lineal en forma matricial.

Supongamos una muestra de tamaño n. Para cada observación i de esta muestra el modelo verifica:

$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \cdots + \beta_k x_{k_i} + \varepsilon_i \quad i = 1, \dots, n$$

De esta manera, podemos definir las siguientes matrices:

$$\vec{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \vec{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} x_{1_1} & \cdots & x_{k_1} \\ \vdots & \ddots & \vdots \\ x_{1_n} & \cdots & x_{k_n} \end{pmatrix} \quad \vec{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Y escribir el modelo matricial:

$$\vec{\mathbf{Y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$$

El método más utilizado para la estimación de los coeficientes del modelo, es el método de mínimos cuadrados ordinarios (MCO). Este método puede utilizarse bajo las siguientes hipótesis a priori:

Sobre la especificación del modelo:

La especificación de la forma funcional es la adecuada.

Los coeficientes $\beta_0, \beta_1, \dots, \beta_k$ son constantes para cambios en la muestra.

Sobre las variables explicativas del modelo:

Son deterministas, es decir, no son variables formadas por valores estimados.

Las variables $X_1, X_2, \dots, X_k$ influyen sobre la variable dependiente Y, y esta relación es unidireccional en el sentido de que el valor de Y no puede influir en los valores tomados por $X_1, X_2, \dots, X_k$.

Ausencia de multicolinealidad grave. Es decir, no existe una relación lineal fuerte entre las variables explicativas consideradas en el modelo.

Sobre las perturbaciones aleatorias:

$$\boldsymbol{\varepsilon} \sim N\left(\vec{\mathbf{0}}_n, \sigma_\varepsilon^2 \mathbf{I}_\mathrm{n}\right)$$

Es decir, los errores del modelo están Normalmente distribuidos con media $\mu_\varepsilon = 0$, y con matriz de covarianzas escalar, lo que implica que la varianza residual σ_ε^2 es constante (hipótesis de homocedasticidad), y que la covarianza entre dos errores cualesquiera es siempre 0, $Cov(\varepsilon_i, \varepsilon_j) \neq 0;\ i, j = 1, 2, \dots, n; i \neq j,$ (hipótesis de ausencia de autocorrelación)

Suponiendo ciertas las hipótesis anteriores, fundamentalmente, las relativas a los residuos, se puede afirmar que:

$$\mathbf{Y} \sim N\left(\mathbf{X}\vec{\boldsymbol{\beta}}, \sigma_\varepsilon^2 \mathbf{I}_n\right)$$

El método MCO nos permite obtener, bajo estas condiciones, los estimadores $\hat{\vec{\boldsymbol{\beta}}}_{\boldsymbol{MCO}}$, que verifican, según el teorema de Gauss-Markov, que son estimadores lineales insesgados y de mínima varianza.

Se basa en minimizar la suma de los cuadrados de los errores del modelo para los valores muestrales dados:

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \vec{\boldsymbol{\varepsilon}}^t \vec{\boldsymbol{\varepsilon}}$$

Como

$$\vec{\boldsymbol{\varepsilon}} = \vec{\mathbf{Y}} - \boldsymbol{X}\vec{\boldsymbol{\beta}}$$

Se verifica que:

$$\vec{\boldsymbol{\varepsilon}}^{\mathbf{t}}\vec{\boldsymbol{\varepsilon}} = \left(\vec{\mathbf{Y}} - \mathbf{X}\vec{\boldsymbol{\beta}}\right)^{\boldsymbol{t}}\left(\vec{\mathbf{Y}} - \mathbf{X}\vec{\boldsymbol{\beta}}\right) = \left(\vec{\mathbf{Y}}^{\boldsymbol{t}} - \vec{\boldsymbol{\beta}}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\right)\left(\vec{\mathbf{Y}} - \mathbf{X}\vec{\boldsymbol{\beta}}\right)$$

$$\vec{\boldsymbol{\varepsilon}}^{\mathbf{t}}\vec{\boldsymbol{\varepsilon}} = \vec{\mathbf{Y}}^{\mathbf{t}}\vec{\mathbf{Y}} + \vec{\boldsymbol{\beta}}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{X}\vec{\boldsymbol{\beta}} - 2 \cdot \vec{\boldsymbol{\beta}}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\vec{\mathbf{Y}}$$

Para minimizar esta función, hallamos el valor de $\vec{\boldsymbol{\beta}}$ para el que la derivada de $\vec{\boldsymbol{\varepsilon}}^{\mathbf{t}}\vec{\boldsymbol{\varepsilon}}$ respecto a $\vec{\boldsymbol{\beta}}$ es 0:

$$\frac{d(\vec{\boldsymbol{\varepsilon}}^{\mathbf{t}}\vec{\boldsymbol{\varepsilon}})}{d\vec{\boldsymbol{\beta}}} = 2\mathbf{X}^{\mathbf{t}}\mathbf{X}\vec{\boldsymbol{\beta}} - 2 \cdot \mathbf{X}^{\mathbf{t}}\vec{\mathbf{Y}} = 0$$

Por lo que bajo el supuesto de que la matriz $\mathbf{X}^{\mathbf{t}}\mathbf{X}$ sea regular, que es equivalente a exigir inexistencia de multicolinealidad entre dos o más variables explicativas (hipótesis ya impuesta) la estimación de los coeficientes del modelo es:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}} = (\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{t}}\vec{\mathbf{Y}}$$

2.4.3. Medidas de bondad del modelo

Antes de comenzar el proceso de validación de un modelo de regresión, es recomendable analizar la existencia de valores atípicos que pudieran influir considerablemente en la estimación de los coeficientes de regresión para posteriormente tomar la decisión sobre la idoneidad de eliminar tales observaciones en caso de que éstas

existan. El estudio de los valores atípicos en la muestra se deja para un capítulo posterior de este trabajo.

Una vez estimados los coeficientes del modelo de regresión lineal, y, por tanto, construido este, y analizada la existencia de valores atípicos, podemos medir la bondad del modelo construido en dos sentidos:

- Bondad de ajuste. Es la capacidad del modelo para estimar los valores dados. Las medidas más utilizadas son el coeficiente de determinación y los criterios de información.
- Capacidad predictiva. Es la capacidad del modelo para estimar correctamente los valores de la variable dependiente de un nuevo conjunto de datos que suele denominarse conjunto de validación, y que suele escogerse aleatoriamente del conjunto original de datos, y reservarse para el cálculo de estas medidas. Todas ellas se basan en el cálculo de los residuos de la estimación de estos valores. Las más utilizadas son: el error absoluto medio, el error relativo medio y el error cuadrático medio.

Para medir la bondad del ajuste realizado, podemos utilizar el coeficiente de determinación, denotado por R^2, y que tiene la siguiente expresión:

$$R^2 = \frac{S_{\hat{Y}}^2}{S_Y^2}$$

El coeficiente de determinación, por tanto, se define como la proporción de varianza de la variable dependiente explicada por el modelo de regresión estimado.

Aplicando el teorema de descomposición de la varianza tenemos esta otra expresión equivalente:

$$R^2 = 1 - \frac{S_{\hat{\varepsilon}}^2}{S_Y^2}$$

Este coeficiente siempre tomará valores entre 0 y 1, salvo cuando el modelo carezca de término independiente.

El inconveniente principal del coeficiente de determinación es que es sensible al número de variables explicativas que tiene el modelo, de manera que un modelo al que se le incluya una variable independiente a Y, aumentará su valor del coeficiente de

determinación por el simple hecho de tener una variable más, aunque ésta sea irrelevante en la estimación. Por ello, es necesario usar un coeficiente, invariante ante distinto número de variables explicativas, que nos permita comparar la bondad de dos modelos que se utilicen para estimar la misma variable dependiente. Este es el coeficiente de determinación ajustado o corregido:

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-k}$$

De esta forma, para medir la bondad de ajuste de un modelo, usaremos el coeficiente de determinación, y para comparar la bondad de ajuste de dos modelos dados, no necesariamente con el mismo número de variables explicativas, usaremos el coeficiente de determinación ajustado.

Akaike, (1977) propuso un criterio para determinar una medida relativa, de la bondad de un modelo, basada en la información de Kullback-Leibler. Dado un conjunto de datos, consideramos la función f, desconocida, que ajusta de manera exacta dicho conjunto en términos de distribución de probabilidad. Si denotamos por g la función hallada para realizar esta estimación, la información de Kullback-Leibler es una medida de la cantidad de información perdida por estimar el conjunto de datos a través de g en lugar de f.

Por tanto, dados dos modelos, será mejor aquél cuya pérdida de información respecto a la dada por f sea menor. De esta forma, el criterio de información propuesto por Akaike, (1977) es:

$$AIC = -2\ln\left(L\left(\hat{\theta}\right)\right) - 2(k+1)$$

Donde $L\left(\hat{\theta}\right)$ es el valor máximo de la función de verosimilitud.

Para el caso de normalidad de los residuos del modelo, o en su defecto, cuando la muestra es suficientemente grande este valor puede ser estimado por:

$$\widehat{AIC} = n(ln(2\pi) + 1) + n\ln\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) + 2(k+2)$$

Con k el número de variables dependientes del modelo.

Es importante destacar que AIC es una medida relativa, por lo que sólo podrá ser utilizado para comparar la bondad del ajuste de dos o más modelos, de manera que aquél con menor valor de AIC será el que mejor ajusta el conjunto de datos dado. No obstante, esto no nos da información sobre la bondad absoluta de éste, por lo que tendremos que basarnos en otras medidas absolutas como el coeficiente de determinación, anteriormente visto.

Este coeficiente penaliza negativamente el mayor número de variables explicativas usadas para la construcción del mismo, por lo que verifica el principio de parsimonia.

El factor de penalización en el aumento del número de variables explicativas es, en este caso, independiente del número de datos e igual a 2. Este factor de penalización puede ser modificado según la necesidad de conseguir un modelo más simple. Incluso, puede hacerse depender este factor del número de observaciones disponibles como es en el caso del denominado criterio de información Bayesiana, BIC, propuesto por Schwarz (1978), y que puede estimarse de la siguiente forma:

$$\widehat{AIC} = n(ln(2\pi) + 1) + n\ln\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) + ln(n)(k+2)$$

Puede observarse que el factor de penalización es $ln(n)$. Aunque el crecimiento de esta función es lento, puede calcularse que, por ejemplo, para una muestra de tamaño 400, el factor de penalización es aproximadamente 6. Para n mayor que 7 el factor de penalización es superior que el dado por el criterio de Akaike, antes visto.

No obstante, y dependiendo de la necesidad de encontrar modelos más simples, se puede definir un criterio de información, para un factor de penalización creado por el investigador, de forma que para un factor p, el criterio de información quedaría:

$$\widehat{AIC}_p = n(ln(2\pi) + 1) + n\ln\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) + \boldsymbol{p}(k+2)$$

Los criterios de información han sido ampliamente utilizados en la literatura en técnicas de selección de modelos respecto a las variables explicativas introducidas en el mismo, de manera que se estiman modelos con diferentes variables explicativas para los

que se calcula el valor del criterio de información, optando por el modelo con menor valor de éste.

Los métodos más utilizados son:

- Hacia delante. En este caso se inicia con la función constante y todos los modelos con una única variable explicativa. Se calculan los valores de AIC. Si el menor valor de AIC se corresponde con el modelo constante, el método finaliza, y el mejor modelo será éste.

  Si por el contrario es uno de los modelos de regresión simple, éste se tomará como el mejor modelo posible en este paso.

  En el siguiente, se estiman todos los modelos resultantes de añadir al modelo seleccionado en el paso anterior, una variable explicativa, para de nuevo comparar los valores de AIC. El proceso sigue hasta que ningún modelo alternativo sea mejor que el seleccionado en el paso anterior o hasta que el modelo contenga todas las variables explicativas.
- Hacia atrás. En este caso se parte del modelo completo con todas las variables explicativas y se compara con todos los modelos posibles resultantes de eliminar una variable. Si el modelo completo es el que menor valor de AIC tenga el procedimiento finaliza, en caso contrario se selecciona el nuevo modelo y se compara con los resultantes de eliminar una nueva variable. El proceso finaliza cuando ningún modelo mejore al previamente seleccionado, o cuando se llegue al modelo constante.
- Paso a paso. En este caso se parte del modelo completo y en cada paso se van eliminando o añadiendo variables anteriormente eliminadas con el mismo criterio.

La bondad de ajuste de un modelo mide la capacidad de éste para ajustarse al conjunto de valores dado. No obstante, una alta bondad de ajuste no implica necesariamente que el modelo tenga capacidad de realizar buenas predicciones dado un nuevo conjunto de datos, que se denomina conjunto de validación.

Para medir la capacidad predictiva de un modelo, la muestra de datos disponible se divide, de forma aleatoria, en dos submuestras: la muestra que se utiliza para estimar el modelo de regresión, y la muestra de validación, que utilizaremos para determinar la capacidad del modelo estimado para estimar valores distintos a los utilizados en el proceso de estimación.

Los errores de estimación del modelo para el nuevo conjunto de datos nos dan información sobre la capacidad predictiva del modelo. Así, para una muestra de

validación $(x_{1i}, x_{2i}, \dots, x_{ki})\ con\ i = 1, \dots, m$ para los que los valores de la variable dependiente son $(y_1, y_2, \dots, y_m)$ podemos calcular:

- Error absoluto medio:

$$EAM = \frac{1}{m}\sum_{i=1}^{m}|\hat{\varepsilon}_i| = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

- Error relativo medio:

$$EAM = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{\varepsilon}_i}{y_i}\right|$$

- Error cuadrático medio:

$$ECM = \frac{1}{m}\sum_{i=1}^{m}\hat{\varepsilon}_i^2$$

2.4.4. Validación del modelo

Una vez estimado el modelo de regresión y analizada la bondad de ajuste y predicción de éste, el siguiente paso será contrastar la veracidad de las hipótesis planteadas anteriormente.

Una vez superada la fase de validación del modelo, podemos utilizarlo para realizar estimaciones de valores de Y a partir de nuevos valores conocidos de $X_1, X_2, \dots, X_k$.

A continuación, en la Tabla 4, se muestran las hipótesis que se desean verificar y el método utilizado para esto.

Tabla 4.Contrastes de validación del modelo de regresión. (Fuente: Elaboración propia)

Hipótesis a contrastar	Contrastes utilizados
Validación global del modelo	- Contraste F de análisis de la varianza
Validación individual	- Contraste t de relevancia individual
Ausencia de multicolinealidad	- Cálculo de los índices VIF. - Cálculo del índice de condición e índice de condición normalizado.
Normalidad de los residuos	- Contraste de Jarque – Bera - Contraste de Shapiro – Wilk - Contraste de Lilliefors (Kolmogorov – Smirnov)
Ausencia de heterocedasticidad	- Contraste de White. - Contraste de Goldfeld – Quandt - Contraste de Breusch – Pagan
Ausencia de autocorrelación	- Contraste de Durbin – Watson - Contraste de Breusch – Godfrey
Especificación del modelo	- Contraste Reset de Ramsey - Contrastes de Davidson y MacKinnon

Fuente: Elaboración propia

2.4.4.1. Análisis de la varianza

El primer contraste a considerar es el análisis de la varianza, para analizar la validez global del modelo en términos poblacionales.

Las hipótesis a contrastar son:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \ ... = \beta_k = 0 \\ H_1: Algún\ \beta_j \neq 0;\ 1 \leq j \leq k \end{cases}$$

La hipótesis nula que se contrasta es muy débil ya que sólo podrá asumirse como cierta en el caso de que todas las variables explicativas sean irrelevantes, y por tanto la fase de especificación del modelo ha fracasado, porque la variable Y no depende de ninguna de las variables consideradas.

El estadístico F se calcula a partir de la Tabla 5:

Tabla 5.Tabla ANOVA regresión.

Fuente de variación	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F
Regresión	$SC_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	$MC_R = \frac{SC_R}{k}$	$F = \frac{MC_R}{MC_E}$
Error	$SC_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-k-1$	$MC_E = \frac{SC_E}{n-k-1}$	
Total	$SC_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$		

Fuente: Elaboración propia

Cuando la hipótesis nula es cierta, F se distribuye según una distribución F de Snedecor con grados de libertad dados por la regresión y el error respectivamente. Así, si se verifica:

$$F > \boldsymbol{\mathcal{F}}(k, n-k-1)_{\boldsymbol{\alpha}}$$

Se rechaza la hipótesis nula, y por tanto podemos afirmar que al menos una de las variables consideradas en la especificación del modelo es relevante para explicar la variable Y.

Habremos incurrido en un error en la especificación del modelo, en la selección de las variables explicativas, en el caso de aceptar la hipótesis nula, y por tanto tendremos que rehacer el planteamiento del problema.

2.4.4.2. Relevancia individual

Una vez aceptada la validez global del modelo de regresión, contrastaremos de forma individual la relevancia o no de cada una de las variables $X_1, X_2, \dots, X_k$. Para ello realizaremos un contraste t, que se obtiene, al igual que el anterior particularizando el teorema de Wald para el contraste de restricciones lineales sobre los parámetros.

En este caso, las hipótesis a contrastar son:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases} \quad \forall j \in \mathbb{N} \,/\, j = 1, .., k$$

Bajo las hipótesis de Normalidad y homocedasticidad de los residuos, el estadístico del contraste verifica:

$$t_j = \frac{\hat{\beta}_j}{\bar{s}_{\hat{\beta}_j}} \sim t(n-k-1)$$

Donde $\bar{s}_{\hat{\beta}_j}$ es la estimación de la desviación típica de β_j. Es fácil observar que si se verifican las hipótesis sobre los residuos:

$$\boldsymbol{\varepsilon} \sim N\left(\vec{\mathbf{0}}_n, \sigma_\varepsilon^2 \mathbf{I}_n\right)$$

De la expresión dada por el método de MCO, el vector de coeficientes $\widehat{\boldsymbol{\beta}}$ también se distribuye Normalmente:

$$\widehat{\vec{\boldsymbol{\beta}}} \sim N\left(\vec{\boldsymbol{\beta}}, \sigma_\varepsilon^2 (\mathbf{X^t X})^{-\mathbf{1}}\right)$$

Por tanto, para aquellas variables para las que sus coeficientes estimados verifiquen:

$$\left|t_j\right| > \; t(n-k-1)_{\alpha/2}$$

Podremos considerarlas relevantes para el modelo.

Es importante destacar que la eliminación de variables de un modelo debe realizarse de manera única en cada paso, y decidir la eliminación de otras variables tras realizar la estimación del modelo en el que se ha omitido dicha variable, ya que la posible existencia de multicolinealidad puede influir en el valor del estadístico de contraste de este test.

Además, la omisión de variables relevantes en el modelo puede sesgar las estimaciones realizadas con este, por lo que la omisión de variables debe realizarse de forma conservadora. Algunos autores, como Caridad (1998), recomiendan incluso aumentar el nivel de significación del contraste más allá del 10%.

2.4.4.3. Multicolinealidad

En los modelos de precios hedónicos es habitual la existencia de numerosas variables explicativas que miden los valores de las distintas características asociadas al bien. Es por ello que hay que tener en cuenta la posible existencia de multicolinealidad grave entre ellas.

La presencia de multicolinealidad grave en un modelo de regresión provoca el aumento de la varianza de los estimadores, lo que provoca inconsistencia de éstos, y la asignación como variables irrelevantes de variables que no lo son.

En numerosos trabajos como en el de Caridad, Tabales, y Ceular (2008) se ha solucionado este problema creando índices que representan un conjunto de características dadas, como índices de calidad o de ubicación.

Otros autores han optado simplemente por eliminar variables irrelevantes del modelo y aquellas que pudieran estar ocasionando problemas de multicolinealidad.

Una alternativa a lo anterior es utilizar otro método de estimación de los parámetros del modelo como puede ser el estimador de Ridge, que consiste en realizar una transformación en la matriz de datos $\mathbf{X^tX}$ que evite la proximidad de su determinante a 0. Esto genera estimaciones sesgadas pero que en ocasiones son más eficientes que las dadas por el método de mínimos cuadrados ordinario.

Existen diferentes métodos para detectar la posible existencia de multicolinealidad grave. Aquí destacamos dos:

Para cada variable $X_1, X_2, \dots, X_k$ construimos el modelo de regresión:

$$X_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_k X_k + \varepsilon$$

A continuación, calculamos los coeficientes de determinación de cada uno de los modelos:

$$R_1^2, R_2^2, \dots, R_k^2$$

Cada coeficiente R_j^2 mide el porcentaje de variabilidad de la variable X_j explicado por el resto de las variables explicativas del modelo, por lo que valores altos de los coeficientes de determinación expresan multicolinealidad grave.

Se define el índice VIF - *variance inflation factor* - para la variable X_j como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

El índice VIF mide cuanto se ha incrementado la varianza de los coeficientes debido a la multicolinealidad.

Valores entre 5 y 10, que es equivalente a valores de R_j^2 entre 0.8 y 0.9, denotan una multicolinealidad moderada.

Valores mayores de 10 indican la existencia de grave multicolinealidad. (Kleinbaum et al., 2013).

El cálculo del índice de condición se basa en la teoría del análisis de componentes principales, en la que las varianzas de las componentes principales son los autovalores de la matriz de varianzas – covarianzas.

Autovalores, de la matriz de datos X, con valores cercanos a 0 muestran presencia de multicolinealidad. Para cuantificar la cercanía de un autovalor a 0, se calcula la raíz cuadrada del cociente entre el mayor de ellos y dicho autovalor.

$$ic(\lambda_i) = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

Por lo que definimos el número de condición como:

$$\kappa(\mathbf{X}) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Aunque es posible calcular el índice de la matriz de datos, es recomendable normalizar la escala de las variables para el cálculo, dividiendo cada una de las columnas de X entre la raíz de la suma de los cuadrados de los valores de ésta. De esta forma se define el número de condición normalizado, más fácilmente interpretable, como:

$$\kappa(\mathbf{X}^*) = \sqrt{\frac{\lambda^*_{max}}{\lambda^*_{min}}}$$

Donde X^* es la matriz de datos normalizada.

Aunque creado por Rachudel (1971), fue desarrollado por Belsley, Kuh, y Welsch (1980), quienes proponen valores superiores a 30 como indicadores de una fuerte multicolinealidad. Valores entre 20 y 30 denotan presencia de una moderada multicolinealidad. Por tanto, podemos descartar la existencia de multicolinealidad que pueda afectar a la estimación de los coeficientes, cuando el número de condición tome valores inferiores a 20.

Existen métodos de regresión sesgada, alternativos al método de mínimos cuadrados ordinario, para los que la varianza de los estimadores es menor, por lo que éstos son más eficientes, cuando se aprecia una fuerte multicolinealidad entre las variables explicativas del modelo.

Entre estos métodos, podemos destacar el creado por Hoerl y Kennard, (1970), conocido como el método de los estimadores Ridge, y que los define como:

$$\widehat{\vec{\boldsymbol{\beta}}}_{\boldsymbol{RIDGE}} = (\mathbf{X^t X} + \lambda \mathbf{I})^{-\mathbf{1}} \mathbf{X^t} \vec{\mathbf{Y}}$$

Esta forma de definir los estimadores del modelo de regresión introduce sesgo en las estimaciones de los parámetros, pero está demostrado que, en estas circunstancias, existe un intervalo de valores de λ para los que el error cuadrático medio de los estimadores es menor que el que se obtiene por el método MCO.

La clave de los estimadores Ridge, radica en hallar el valor de λ que hace mínima la varianza de los estimadores de los coeficientes del modelo. Existen para ello diferentes métodos, entre ellos el de selección del valor de λ mediante el uso de trazas ridge. Como el objetivo es reducir la variabilidad de la estimación de los coeficientes del modelo, se generan las estimaciones de los coeficientes para pequeños incrementos en el valor de λ, de modo que seleccionaremos aquél en el que las estimaciones se estabilizan.

2.4.4.4. Normalidad residual

La última fase de validación de un modelo de regresión es la relativa a los residuos. Como se ha explicado anteriormente, todo lo realizado anteriormente está basado en el cumplimiento de:

$$\vec{\boldsymbol{\varepsilon}} \sim N\left(\vec{\mathbf{0}}_n, \sigma_\varepsilon^2 \mathbf{I}_n\right)$$

El estudio de la normalidad de los residuos del modelo puede iniciarse por una representación gráfica que aporta información preliminar sobre el posible resultado del contraste. Entre los distintos gráficos que pueden realizarse podemos destacar dos: el histograma y el gráfico QQ. Con el primero de ellos se trata de analizar la similitud entre la función de densidad empírica y la teórica, analizando las similitudes de la representación gráfica de la primera, con la campana de Gauss.

El gráfico QQ es un gráfico de dispersión en el que las coordenadas de sus puntos vienen definidas por los cuantiles de la distribución empírica y sus correspondientes en la distribución normal. La distribución de residuos se asemejará a la distribución Normal cuanto más alineados estén los puntos del gráfico a la bisectriz del primer cuadrante.

Una vez realizado un análisis previo del cumplimiento de la hipótesis de normalidad, podemos realizar un contraste de hipótesis. Existen numerosos test para determinar la bondad de ajuste a la distribución Normal, debido a la importancia que esta tiene en el proceso inferencial. Entre estos métodos destacaremos tres: El método de Jarque – Bera, el de Shapiro – Wilks, y el de Kolmogorov – Smirnov.

En primer lugar, analizaremos la normalidad de los residuos a través del contraste de Jarque - Bera. Este contraste fue propuesto por Jarque y Bera (1980). Se basa en los coeficientes de Fisher que miden la simetría y el apuntamiento de la distribución:

$$g_1 = \frac{m_3}{s^3} \qquad\qquad g_2 = \frac{m_4}{s^4}$$

Donde s representa la desviación típica, y m_k los momentos centrales de orden k de la variable. Para una variable X formada por los valores $x_1, x_2, \ldots, x_n$ se define como:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Las hipótesis del contraste de Normalidad de Jarque – Bera para los residuos son:

$$\begin{cases} H_0: \varepsilon \sim Normal \\ H_1: \varepsilon \nsim Normal \end{cases}$$

Los autores demuestran que el estadístico de contraste, bajo hipótesis nula, sigue una distribución Chi – cuadrado con 2 grados de libertad:

$$JB = n\left[\frac{m_3^2}{6 \cdot m_2^3} + \frac{1}{24}\left(\frac{m_4}{m_2^2} - 3\right)^2\right] + n\left[\frac{3m_1^2}{2m_2} - \frac{m_3 m_1}{m_2^2}\right] \sim \chi^2(2)$$

O equivalentemente:

$$JB = n\left(\frac{g_1^2}{6} + \frac{(g_2 - 3)^2}{24}\right) \sim \chi^2(2)$$

Por tanto, no existen evidencias significativas para rechazar la normalidad de los residuos si:

$$JB < \chi^2(2)_\alpha$$

Otro contraste de normalidad ampliamente utilizado es el propuesto por Shapiro y Wilk (1965). En su formulación inicial sólo era válido para muestras con tamaño inferior a 50. En este caso es además una prueba más potente que la próxima de Kolmogorv – Smirnov que explicaremos a continuación.

No obstante, tras diversas actualizaciones del contraste, implementadas en las funciones de R utilizadas habitualmente, su utilidad se ha ampliado a muestras de cualquier tamaño según explican Razali y Wah (2011)

Las hipótesis del contraste, al igual que en el caso anterior son:

$$\begin{cases} H_0: \varepsilon \sim Normal \\ H_1: \varepsilon \nsim Normal \end{cases}$$

El estadístico de contraste es:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \varepsilon_{(i)}\right)^2}{\sum_{i=1}^{n} \varepsilon_i^2}$$

Donde $\varepsilon_{(1)}, \varepsilon_{(2)}, \dots, \varepsilon_{(n)}$ son los residuos del modelo ordenados.

El vector $(a_1, a_2, \dots, a_n)$ se calcula como:

$$\vec{\boldsymbol{a}}^{\boldsymbol{t}} = (a_1, a_2, \dots, a_n) = \frac{\vec{\mathbf{m}}^{\mathbf{t}}\mathbf{V}^{-\mathbf{1}}}{(\vec{\mathbf{m}}^{\mathbf{t}}\mathbf{V}^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}}\vec{\mathbf{m}})^{1/2}}$$

Con $\vec{\mathbf{m}}^{\boldsymbol{t}} = (m_1, m_2, \dots, m_n)$ el vector de valores esperados de los estadísticos ordenados de la distribución Normal estándar, y $\mathbf{V}$ su correspondiente matriz de covarianzas de orden n.

Definido de esta forma el estadístico de contraste, se demuestra en Shapiro y Wilk, (1965), que este siempre toma valores menores que 1, de forma que valores pequeños nos ofrecen indicios de la ausencia de normalidad en la distribución de los datos.

Los valores de probabilidad límite fueron tabulados, para distintos valores del estadístico y del tamaño muestral hasta $n = 50$ por los autores. No obstante, según las características de $(a_1, a_2, \dots, a_n)$, para tamaños de muestra entre 12 y 5000, se demuestra que $ln(1 - W)$ sigue una distribución aproximadamente Normal.

Otro contraste, basado en los trabajos de Kolmogorov, (1933) y Smirnov, (1948) y desarrollado por Massey Jr, (1951) es el conocido como contraste de Kolmogorov – Smirnov, y permite analizar la bondad de ajuste de un conjunto de datos a una determinada distribución teórica, no sólo la distribución Normal.

Para la realización del contraste, previamente se ordenarán los residuos de menor a mayor, de forma que:

$$\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \dots \leq \varepsilon_{(n)}$$

Y para estos valores, se construye su función de distribución, definida para cualquier valor k, como $S_n(x) = {}^{k}/_{n}$.

Sea $F_0(x)$ la función de distribución de la distribución teórica que se desea contrastar, en este caso la distribución Normal.

Entonces, diremos que el conjunto de datos estudiado sigue una distribución Normal, si su función de distribución empírica así definida coincide con la función de distribución teórica en esos puntos. De ahí que las hipótesis de contraste sean:

$$\begin{cases} H_0: S_n(x) \equiv F_0(x) \\ H_1: S_n(x) \not\equiv F_0(x) \end{cases}$$

El estadístico de contraste se basa en calcular la mayor diferencia entre la función de probabilidad acumulada teórica y la empírica. Por lo que se define como:

$$D = \max_{1 \le i \le n} |S_n(\varepsilon_i) - F_0(\varepsilon_i)|$$

Un valor grande del estadístico mostrará gran discrepancia entre las funciones de distribución, por lo que nos permitirá rechazar la hipótesis nula y por tanto la normalidad de los residuos. Los valores críticos del estadístico son tabulados y dependen de la distribución teórica a la que se desee aproximar.

Años más tarde, Lilliefors (1967), proporcionó una nueva tabla de valores críticos del contraste de ajuste a una distribución Normal, para el caso en el caso es necesario estimar los parámetros de la distribución a través de la muestra por ser éstos desconocidos.

Según el análisis comparativo de los test de normalidad realizado por Razali y Wah, (2011), en el que no se incluyó el test de Jarque – Bera, en general, la potencia del test de Shapiro – Wilk es ligeramente superior a la de Lilliefors, y por supuesto muy superior a la de Kolmogorov – Smirnov, cuando se estudia ésta en función del tamaño muestral.

2.4.4.5. Homocedasticidad residual

Para estudiar el cumplimiento de la hipótesis de homocedasticidad podemos existen varios contrastes de hipótesis posibles. En este trabajo destacaremos tres: la prueba de Goldfeld – Quandt, el test de White y el de Breusch – Pagan.

El primero de ellos es presentado por Goldfeld y Quandt (1965) junto a un contraste no paramétrico para el estudio de la homocedasticidad de los residuos de un modelo de regresión.

Es recomendable comenzar el estudio de la heterocedasticidad a través de una representación gráfica. Diferentes gráficos de dispersión pueden ofrecer información relevante sobre la heterocedasticidad de los residuos del modelo y sus causas.

El gráfico de residuos frente a la variable predicción $\hat{y}$ nos puede informar no sólo de presencia de heterocedasticidad, sino también de autocorrelación o de no linealidad. La ordenación de las predicciones en sentido creciente y su representación junto a los residuos puede revelarnos presencia de heterocedasticidad al observar comportamientos difícilmente explicables por el azar.

Un gráfico de dispersión que relacione los residuos, o los residuos cuadráticos, con cualquier variable incluida en el modelo puede indicarnos qué variable o variables causan el problema.

Para ello basta con ordenar los residuos en orden creciente de cada una de las variables explicativas, y representar el gráfico $X_k - \hat{\varepsilon}$. En caso de existir heterocedasticidad, ésta puede ser creciente o decreciente como puede observarse en los ejemplos extraídos del libro de Econometría de Caridad (1998), correspondientes a la

Figura *1* y la Figura 2.

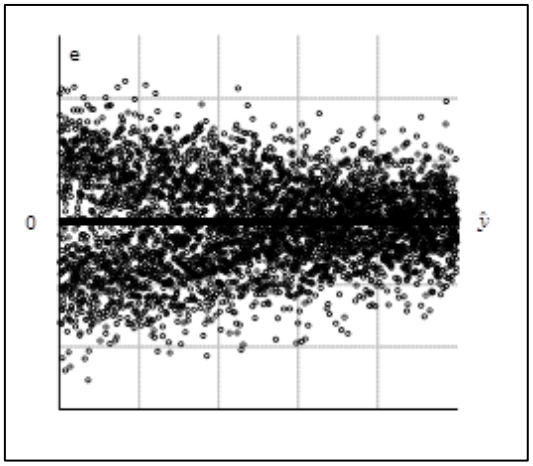


Figura 1. Heterocedasticidad decreciente. Fuente: Caridad (1998)

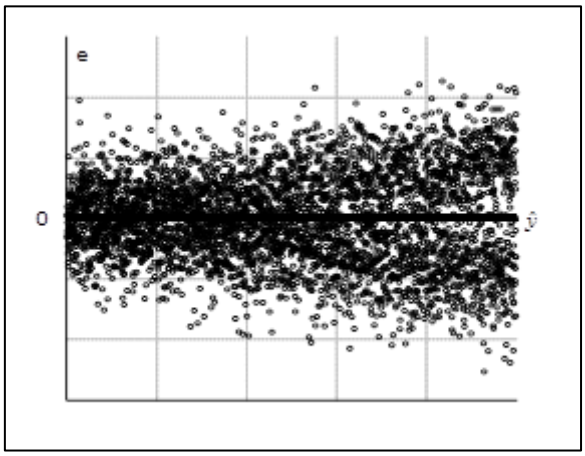


Figura 2. Heterocedasticidad decreciente. Fuente: Caridad, (1998)

Una vez analizada la posible existencia de heterocedasticidad, mediante representaciones gráficas, respecto a la variable X_k, y tras reordenar los datos en sentido creciente respecto a esta variable, podemos aplicar el contraste de Goldfeld – Quandt. Para ello, se divide la muestra en tres grupos de forma creciente según los valores de X_k de la siguiente forma:

$$x_{k_1}, x_{k_2}, \dots, x_{k_{\frac{n-p}{2}}}, x_{k_{\frac{n-p}{2}+1}}, \dots, x_{k_{\frac{n+p}{2}}}, x_{k_{\frac{n+p}{2}+1}}, \dots, x_{k_n}$$

De forma que la primera y la tercera submuestra, que contienen los valores más pequeños y más grandes de X_k, contengan $\frac{n-p}{2}$ datos y la submuestra central contenga p datos.

Una vez realizada esta partición de la muestra, se toman las submuestras primera y tercera, se estiman dos modelos de regresión a partir de estos datos y se calculan las sumas de los cuadrados de los errores de cada modelo que denotaremos S_1 y S_3 respectivamente. La existencia de grandes diferencias entre ambas denota existencia de heterocedasticidad residual dependiente de la variable X_k.

Las hipótesis a contrastar son:

$$\begin{cases} H_0: Var(\varepsilon) = \sigma_\varepsilon^2 \\ H_1: Var(\varepsilon) = f(X_k) \end{cases}$$

Como se ha indicado anteriormente, el estadístico a contrastar es el cociente de las sumas de los cuadrados de los modelos, que bajo hipótesis nula sigue una distribución F – Snedecor:

$$G = \frac{S_1}{S_3} \sim \boldsymbol{\mathcal{F}}\left(\frac{n-p}{2} - k - 1, \frac{n-p}{2} - k - 1\right)$$

Por tanto, si se verifica que:

$$G > \boldsymbol{\mathcal{F}}\left(\frac{n-p}{2} - k - 1, \frac{n-p}{2} - k - 1\right)_{\boldsymbol{\alpha}}$$

Concluiremos la existencia de heterocedasticidad de los residuos.

La idea que subyace de descartar los p valores centrales de la muestra ordenada es aumentar la potencia del contraste al hacer más claras las posibles diferencias que puedan existir en la varianza residual de cada grupo.

La selección del valor de p es un tema discutido por diferentes autores, debido a que cuanto mayor es su valor, mayores serán las diferencias entre los valores de ambas muestras. No obstante, un elevado valor de p reduce los grados de libertad del estadístico calculado.

El criterio más usado es el de dividir la muestra en tres submuestras aproximadamente iguales en tamaño, obligando a que las submuestras 1 y 3 contengan el mismo número de datos.

Una prueba alternativa para la detección de heterocedasticidad es el test de White. Al contrario que en el contraste visto anteriormente, éste no requiere de la hipótesis de normalidad de los residuos. Además, tampoco es necesario detectar previamente la variable o variables explicativas causantes de la heterocedasticidad, ya que es de tipo genérico.

Este contraste fue propuesto por Halbert White en 1980, y se basa en contrastar la validez de un modelo de regresión que explique el cuadrado de los residuos del modelo original a partir de las variables explicativas, sus cuadrados y sus pares de productos cruzados.

El modelo auxiliar con el que analiza la existencia de heterocedasticidad es, por tanto:

$$\hat{\varepsilon}^2 = \alpha_0 + \sum_{i=1}^{k} \alpha_i \cdot X_i + \sum_{i=1}^{k} \sum_{j=i}^{k} \alpha_{ij} \cdot X_i \cdot X_j + \varepsilon'$$

Si el modelo es globalmente válido, la varianza residual puede ser explicada a partir de las variables explicativas del modelo, sus cuadrados y sus productos cruzados, por lo tanto, las perturbaciones aleatorias son heterocedásticas.

Una forma de analizar la presencia de heterocedasticidad residual es realizar contraste de validación global del modelo auxiliar, y determinar su existencia si el modelo es globalmente válido.

No obstante, el estadístico de White, cuya expresión se debe a Breusch y Pagan, es:

$$W = n \cdot R'^2$$

Con n el tamaño muestral y R'^2 el coeficiente de determinación del modelo auxiliar anterior. Este estadístico se distribuye según una Chi – cuadrado con p grados de libertad, donde p es el número de variables que contiene dicho modelo, es decir,

$$p = k + k + \frac{k \cdot (k-1)}{2} = \frac{k^2 + 3k}{2}$$

Por lo que, si se verifica que:

$$W = n \cdot R'^2 > \chi^2 \left(\frac{k^2 + 3k}{2} \right)_{\alpha}$$

Podremos afirmar la existencia de heterocedasticidad residual.

Este método es más general que el anterior ya que no requiere del cumplimiento de algunas hipótesis como por ejemplo la normalidad de los residuos. Sin embargo, al contrario que con el método anterior, si se confirma la existencia de heterocedasticidad, el test de White no indica la variable o variables que la causan.

No obstante, la imposibilidad de construir el modelo auxiliar del contraste en presencia de variables *dummies* limita los casos en los que éste es aplicable, debido a que las distintas potencias de una variable de este tipo son iguales a la variable original, y esto genera un problema de multicolinealidad exacta.

Otro contraste para el estudio de la heterocedasticidad residual es el propuesto por Breusch y Pagan (1979). Este contraste tiene características de los dos contrastes anteriores. Por un lado, sirve para detectar la heterocedasticidad generada por un conjunto específico de variables como el de Goldfeld – Quandt, y por otro, utiliza la metodología basada en la construcción de un modelo de estimación de la varianza residual para la realización del contraste.

Las hipótesis propuestas por los autores son las siguientes:

$$\begin{cases} H_0: Var(\varepsilon) = \sigma_\varepsilon^2 \\ H_1: Var(\varepsilon) = h(\boldsymbol{\alpha}, \vec{\boldsymbol{Z}}) \end{cases}$$

Donde $\vec{\boldsymbol{Z}}$ es el vector de variables de las que puede depender la varianza residual, y $\boldsymbol{\alpha}$ es el vector de coeficientes del modelo que las relaciona, de forma que:

$$\sigma_\varepsilon^2 = \alpha_0 + \alpha_1 \cdot Z_1 + \cdots + \alpha_p \cdot Z_p + \varepsilon^*$$

Por tanto, la formulación de las hipótesis es equivalente a:

$$\begin{cases} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0 \\ H_1: \exists i \ / \ \alpha_i \neq 0 \ con \ con \ 1 \leq i \leq p \end{cases}$$

Para la aplicación del contraste, se calculan los residuos del modelo: $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n$, sus cuadrados, y éstos se estandarizan dividiéndolos por el error cuadrático medio, es decir, para todo $1 \leq i \leq n$:

$$\tilde{\varepsilon}_i^2 = \frac{\hat{\varepsilon}_i^2}{\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

Para los errores cuadráticos estandarizados se construye el modelo:

$$\tilde{\varepsilon}_i^2 = \alpha_0 + \alpha_1 Z_1 + \cdots + \alpha_p Z_p + \varepsilon^*$$

Y para éste, su coeficiente de determinación: $R^2_{\tilde{\varepsilon}^2}$, ya que, si el porcentaje de variabilidad explicada por el modelo es alto, significa que alguna de la estimación de la varianza residual depende de alguna de las variables seleccionadas, y por tanto los residuos del modelo presentan heterocedasticidad.

Equivalentemente, si la varianza residual del modelo estimado es grande, la capacidad de ajuste del modelo es pequeña y por tanto aceptaremos la hipótesis nula de ausencia de heterocedasticidad. Por tanto, el estadístico de contraste puede definirse como:

$$W = \frac{n}{2} S^2_{\tilde{\varepsilon}^2}$$

Que sigue una distribución Chi cuadrado con p grados de libertad. Luego se aceptará la hipótesis de homocedasticidad si se verifica:

$$W = \frac{n}{2} S^2_{\tilde{\varepsilon}^2} > \chi^2(p)_\alpha$$

El incumplimiento de la hipótesis de homocedasticidad conlleva la pérdida de la eficiencia de los estimadores de los coeficientes del modelo. Las varianzas de éstos también están erróneamente estimadas, lo que tiene como consecuencia que los estadísticos t de relevancia no sean fiables.

Para resolver la existencia de heterocedasticidad en un modelo de regresión pueden tomarse medidas como aplicar transformaciones logarítmicas sobre las variables,

aumentar el tamaño muestral o definir variables que tomen valores relativos como índices porcentuales. Estas estrategias no siempre dan resultado.

Una alternativa a todo lo anterior es aplicar el método de mínimos cuadrados generalizado (MCG). Este método consiste en realizar una transformación sobre los datos, denominada transformación de Aitken, que haga que la matriz de covarianzas de los residuos sea una matriz escalar.

A partir del modelo de regresión en forma matricial:

$$\vec{\mathbf{Y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$$

El objetivo es hallar una matriz $\boldsymbol{H}$, de forma que al aplicar la transformación:

$$\mathbf{H}\vec{\mathbf{Y}} = \mathbf{H}\mathbf{X}\vec{\boldsymbol{\beta}} + \mathbf{H}\vec{\boldsymbol{\varepsilon}}$$

Se cumpla que:

$$Var(\mathbf{H}\vec{\boldsymbol{\varepsilon}}) = \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\mathbf{I}_{\boldsymbol{n}}$$

Si un modelo presenta heterocedasticidad y autocorrelación, su matriz de covarianzas será de la forma:

$$Var(\vec{\boldsymbol{\varepsilon}}) = \sigma_{\varepsilon}^{2}\boldsymbol{\Sigma}_{\boldsymbol{n}}$$

Tras la transformación de Aitken, la matriz de covarianzas verifica:

$$Var(\mathbf{H}\vec{\boldsymbol{\varepsilon}}) = \sigma_{\varepsilon}^{2}\mathbf{H}^{\mathbf{t}}\boldsymbol{\Sigma}_{\mathrm{n}}\mathbf{H}$$

Para que esta matriz sea diagonal, la matriz **H** es tal que:

$$\mathbf{H}^{\mathbf{t}}\boldsymbol{\Sigma}_{\mathrm{n}}\mathbf{H} = \mathbf{I}_{\boldsymbol{n}}$$

La existencia de esta matriz está basada en la simetría de la matriz de covarianzas. No obstante, la matriz H es desconocida y deberá ser estimada.

En el caso de que la matriz de covarianzas sea diagonal no escalar, es decir, que el modelo presente únicamente heterocedasticidad, la matriz de la transformación de Aitken será de la forma:

$$\mathbf{H} = \begin{pmatrix} \frac{1}{\sigma_{\varepsilon_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{\varepsilon_2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma_{\varepsilon_n}} \end{pmatrix}$$

Ya que, de este modo, es fácil observar que:

$$\boldsymbol{Var}(\mathbf{H}\vec{\boldsymbol{\varepsilon}}) = \boldsymbol{H}^t Var(\mathbf{H}\vec{\boldsymbol{\varepsilon}}) \cdot \mathbf{H} = \sigma_{\varepsilon}^2 \mathbf{I}_{\boldsymbol{n}}$$

Los valores de $\sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \dots, \sigma_{\varepsilon_n}$ son desconocidos, por lo que es necesario estimar su valor. Esto puede hacerse directamente, si para el caso concreto de estudio puede deducirse una forma funcional que estime los valores de las varianzas residuales, por ejemplo, en el caso de confirmar la dependencia de la varianza de una o más de las variables explicativas del modelo, tras haber realizado el test de Goldfeld – Quandt, o el de Breusch – Pagan.

No obstante, si se desconoce la forma funcional adecuada para la estimación de las varianzas residuales puede usarse el método propuesto en White (1980). El autor construye estimadores consistentes de $\sigma_{\varepsilon_i}^2,\ i = 1, \dots, n$ en presencia de heterocedasticidad asignando a estos los valores estimadores para los residuos al cuadrado en el modelo auxiliar anteriormente comentado:

$$\hat{\sigma}_{\varepsilon_i}^2 = \hat{\varepsilon}_i^2 \quad \forall i = 1, \dots, n$$

Este método carece de buenas propiedades cuando del tamaño muestral es reducido, por lo que se desaconseja en este caso.

Como hemos visto, este método corrige únicamente la estimación de las varianzas de los residuos y por tanto la varianza de los estimadores. No obstante, la estimación de los coeficientes queda invariante a través de este método.

2.4.4.6. Ausencia de autocorrelación residual

Si las perturbaciones aleatorias de un modelo de regresión están linealmente relacionadas entre sí, decimos que existe autocorrelación. Normalmente, este problema está asociado a datos temporales ya que es habitual en este tipo de datos que el error en un instante t, ε_t, esté influido por el error del modelo en instantes anteriores.

En los modelos hedónicos, cada observación se corresponde con un bien concreto, luego los datos serán de corte transversal, lo que hace improbable la existencia de autocorrelación residual.

No obstante, los inmuebles son bienes para los que, a priori, su valor está fuertemente relacionado con la ubicación en la que se encuentren. Es por ello, que en este caso tiene más sentido analizar la existencia de autocorrelación espacial a través de la posición de las observaciones en el plano.

Debido a que ésta es una revisión de los modelos de regresión en general en general, estudiaremos a continuación dos contrastes para la detección de autocorrelación y posteriormente el método de estimación de los parámetros más adecuado en caso de existencia.

El contraste de Durbin y Watson, fue presentado por los autores que dan nombre a dicho contraste en (1950). Este contraste permite analizar la existencia de autocorrelación de primer orden, en una serie temporal, es decir, si existe correlación lineal entre el residuo del modelo en el instante t y el del instante inmediatamente anterior. Por tanto, no analiza la existencia de autocorrelación en intervalos de tiempo más amplios.

No obstante, es un contraste muy utilizado debido a que en la práctica los residuos están más fuertemente relacionados cuanto menor es el período de tiempo transcurrido entre ellos. Por tanto, el test detecta la existencia de autocorrelación, pero no el orden de la misma.

En el caso de que exista autocorrelación de primer orden, ésta puede ser modelizada de la siguiente forma:

$$\varepsilon_t = \phi\varepsilon_{t-1} + a_t \quad con - 1 \leq \phi \leq 1$$

Con a_t definido como ruido blanco. Por tanto, si $\phi = 0$ descartaremos la existencia de autocorrelación de primer orden. En caso de que $\phi \neq 0$, diremos que los residuos del modelo siguen un modelo autorregresivo de orden 1 - $AR(1)$ -, y, por tanto, el modelo de regresión incumple la hipótesis relativa a la ausencia de autocorrelación residual.

El estadístico de contraste del test de Durbin y Watson evalúa las diferencias cuadráticas de los residuos en un período determinado y el período inmediatamente anterior, respecto al error cuadrático medio:

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}$$

Si la muestra es suficientemente grande, se verifica que:

$$DW \cong 2\left(1 - \hat{\phi}\right)$$

Por lo que el estadístico DW varía aproximadamente entre 0 y 4. En ese caso, se cumple que:

- Valores de DW cercanos a 2, nos indican valores de $\hat{\phi}$ próximos a 0, y por tanto ausencia de autocorrelación de primer orden.
- Valores de DW cercanos a 0, nos indican valores de $\hat{\phi}$ próximos a 1, y por tanto existencia de autocorrelación positiva de primer orden.
- Valores de DW cercanos a 4, nos indican valores de $\hat{\phi}$ próximos a -1, y por tanto existencia de autocorrelación negativa de primer orden, y por tanto alternancia temporal en el signo de los errores.

El estadístico DW no es sencillo de modelizar ya que depende de cada modelo estimado, y por tanto no se puede calcular un valor crítico para la decisión del contraste. Durbin y Watson (1950) resolvieron el problema asociado a la distribución del estadístico definiendo un intervalo "crítico" de valores (d_L, d_U) dependiente del tamaño muestral y del número de variables explicativas. Valores del estadístico pertenecientes al intervalo significan que la existencia de autocorrelación de primer orden no es concluyente. Los autores construyeron los intervalos críticos de decisión para tamaños muestrales entre 15 y 100, entre 1 y 5 variables explicativas y siempre que el modelo estimado contenga

término independiente. Años más tarde, Savin y White, (1977), ampliaron el cálculo del intervalo para muestras de hasta 200 individuos y hasta 20 variables explicativas.

Por tanto, el contraste de Durbin – Watson se plantea de la siguiente forma. Si el valor de DW es inferior a 2 analizaremos la existencia de autocorrelación positiva de primer orden y la hipótesis a plantear es:

$$\begin{cases} H_0: \phi = 0 \\ H_1: \phi > 0 \end{cases}$$

- Si $DW < d_L$ podemos afirmar, al nivel de significación al que se ha estimado el intervalo crítico, que existe autocorrelación positiva de primer orden.
- Si $DW > d_U$ no existen evidencias de existencia de autocorrelación de primer orden.
- Si $d_L < DW < d_U$ el resultado del contraste no es concluyente.

En otro caso, si DW es superior a 2, analizaremos la existencia de autocorrelación negativa de primer orden:

$$\begin{cases} H_0: \phi = 0 \\ H_1: \phi < 0 \end{cases}$$

- Si $DW > 4 - d_L$ podemos afirmar, al nivel de significación al que se ha estimado el intervalo crítico, que existe autocorrelación negativa de primer orden.
- Si $DW < 4 - d_U$ no existen evidencias de existencia de autocorrelación de primer orden.
- Si $4 - d_U < DW < 4 - d_L$ el resultado del contraste no es concluyente.

Este contraste es muy utilizado y por tanto ampliamente aceptado. No obstante, presenta dos inconvenientes a tener en cuenta:

- En primer lugar, este contraste sólo analiza la existencia de autocorrelación de primer orden. Por tanto, en el que caso de afirmar su existencia no nos indica el posible orden de ésta. Además, en ocasiones, en procesos estacionales es útil analizar la existencia de autocorrelación de órdenes distintos al primero. Por ejemplo, en series mensuales es interesante analizar la existencia de autocorrelación de orden 12. Esto último está en parte resuelto con contrastes como el de Wallis en el que se analiza la existencia de autocorrelación de orden 4, o el de King que lo hace para orden 12.
- En segundo lugar, este contraste no puede ser aplicado en modelos dinámicos, aquellos en los que se consideran variables dependientes retardadas dentro del

grupo de variables explicativas, ya que requiere que la matriz de datos sea determinista. Años más tarde, Durbin (1970) propuso un contraste aplicable a modelos dinámicos utilizando la estimación de la varianza del coeficiente de la regresión de la variable retardada en el modelo.

Otro contraste muy utilizado para la detección de autocorrelación es el propuesto y revisado por Breusch y Godfrey (1981). Mejora en potencia al contraste de Durbin – Watson en potencia y aplicabilidad, ya que se puede aplicar para contrastar estructuras de autocorrelación de cualquier orden, e incluso para estructuras tanto autorregresivas como de medias móviles.

El contraste se basa en la construcción de un modelo de regresión en el que la variable dependiente es la formada por los residuos del modelo original, y como variables independientes tomaremos las variables originales del modelo a la que incluiremos la estructura de autocorrelación de los residuos, es decir:

$$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 X_{1t} + \cdots + \alpha_k X_{kt} + \rho_1 \hat{\varepsilon}_{t-1} + \cdots + \rho_p \hat{\varepsilon}_{t-p} + a_t$$

Para contrastar una estructura autorregresiva de orden p en los residuos.

Partiendo de la hipótesis de independencia de los residuos y las variables independientes por el cumplimiento de la hipótesis de homocedasticidad, en el supuesto de que no exista autocorrelación residual, el modelo anteriormente expuesto no tendrá capacidad para explicar la variabilidad de los residuos del modelo, y, por tanto, su coeficiente de determinación será cercano a 0.

El estadístico de contraste se define como:

$$W = n^* R^2$$

Donde n^* es el tamaño muestral disponible tras la construcción de los residuos retardados, es decir, $n^* = n - p$. El estadístico así definido sigue una distribución Chi cuadrado con p grados de libertad, por lo que para las hipótesis:

$$\begin{cases} H_0: \rho_k = 0 \ \forall k = 1, \dots, p \\ H_1: \exists j \ / \ \rho_j \neq 0 \ j = 1, \dots, p, \end{cases}$$

Afirmamos la existencia de autocorrelación en los residuos si:

$$W = n^* R^2 > \chi^2(p)_\alpha$$

En el caso de que el modelo presente autocorrelación, el estimador hallado para $\vec{\boldsymbol{\beta}}$ por el método de mínimos cuadrados no es eficiente, y la alternativa, al igual que en el caso de existencia de heterocedasticidad residual, una alternativa a este método es el método de mínimos cuadrados generalizado, en el que el objetivo es determinar la matriz **H** de la transformación de Aitken que transforme la matriz de covarianzas residual, en una matriz escalar.

Para el cálculo de esta matriz, es necesario modelizar los residuos a partir de sus retardos y los del ruido blanco y posteriormente considerar este modelo para realizar estimación del modelo de regresión principal.

2.4.5. Forma funcional del modelo

Como se ha comentado anteriormente, la especificación de la forma funcional a considerar al estimar el modelo de regresión dependerá de la información previa que se tenga de las variables a analizar. No obstante, el modelo de regresión lineal es el más utilizado ya que en numerosas ocasiones los modelos no lineales que se aplican son linealizables utilizando transformaciones logarítmicas de las variables en estudio.

Estas transformaciones, nos devuelven aproximaciones a los estimadores del modelo no lineal, pero reducen sustancialmente la complejidad del problema, que, al ser no lineal, tiene un coste computacional NP complejo.

La selección de una forma funcional incorrecta provoca que las estimaciones obtenidas de los parámetros sean sesgadas y poco precisas. Por este motivo, es necesario confirmar si el modelo seleccionado es el más adecuado.

El contraste Reset de Ramsey (1976) permite decidir sobre la idoneidad de utilizar un modelo lineal.

Para ello, una vez estimado el modelo lineal propuesto por el método de mínimos cuadrados ordinario:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Se calcula la variable $\hat{Y}$, formada por las estimaciones dadas por este modelo para la variable dependiente, y la suma de cuadrados de los errores: SC_{E_1}

A continuación, se estima el modelo de regresión lineal al que se le añaden como variables explicativas, diferentes potencias de la estimación de la variable endógena. Esto es:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \sum_{j=2}^{m} \delta_j \hat{Y}^j + \varepsilon$$

Y calculamos, también para este modelo, la suma de los cuadrados de los errores: SC_{E_2}

El modelo lineal será la especificación correcta siempre que se verifique la condición:

$$\delta_j = 0 \quad \forall j = 2, \ldots, m$$

Por tanto, el test F de validez global para un subconjunto de variables del modelo nos dará respuesta a la verificación o no de la condición anterior. Formulamos las hipótesis:

$$\begin{cases} H_0: \delta_2 = \cdots = \delta_m = 0 \\ H_1: \delta_j \neq 0 \; para \; algún \; j = 2, \ldots, m \end{cases}$$

El estadístico de contraste es

$$F = \frac{\frac{\left(SC_{E_1} - SC_{E_2}\right)}{m-1}}{\frac{SC_{E_2}}{n-k-1}}$$

Que bajo hipótesis nula sigue una distribución F – Snedecor, por lo que si

$$F > \boldsymbol{\mathcal{F}}(m-1, n-k-1)_{\boldsymbol{\alpha}}$$

Rechazaremos la hipótesis nula y por tanto la linealidad del modelo de regresión. En este caso, la especificación del modelo lineal no es la más adecuada.

Davidson y MacKinnon proponen en 1981, varios test para contrastar la especificación de un modelo econométrico en presencia de varios modelos posibles que puedan explicar un mismo fenómeno. Si queremos estimar una variable Y, y para ello proponemos dos formas funcionales $f\left(X_i, \vec{\boldsymbol{\beta}}\right)$ (hipótesis nula) y $g\left(Z_j, \vec{\boldsymbol{\lambda}}\right)$ (hipótesis alternativa) con $i = 1, \dots, k$ y $j = 1, \dots, p$, construiremos el modelo auxiliar:

$$Y = (1-\alpha) f\left(X_i, \vec{\boldsymbol{\beta}}\right) + \alpha \hat{g} + \varepsilon$$

Sobre este modelo, Davidson y Mackinnon proponen aplicar el que denominan test J por ser un modelo en el que se estiman conjuntamente (*jointly*) los parámetros α y β. Este modelo coincide con el modelo dado en H_0 si $\alpha = 0$, y con el dado por H_1 si $\alpha = 1$. Los autores recomiendan aplicar este método cuando $f\left(X_i, \vec{\boldsymbol{\beta}}\right)$ es lineal, debido a su sencillez.

Los pasos para la aplicación de este contraste son los siguientes:

- Se estima el modelo de regresión:

$$Y = g\left(Z_j, \vec{\boldsymbol{\lambda}}\right) + \varepsilon$$

- Se calculan los valores de $\hat{Y}$ en el modelo anterior, y se estima el modelo:

$$Y = (1-\alpha) f\left(X_i, \vec{\boldsymbol{\beta}}\right) + \alpha \hat{\vec{\mathbf{g}}} + \varepsilon$$

- Una vez estimado este modelo, se realiza un contraste t de significación individual para el coeficiente α. Si afirmamos que $\alpha \neq 0$, rechazamos la hipótesis nula y por tanto la especificación del modelo de regresión dado por $f\left(X_i, \vec{\boldsymbol{\beta}}\right)$.

Por otro lado, podemos, también, plantear el modelo:

$$Y = (1-\alpha) f\left(X_i, \hat{\vec{\boldsymbol{\beta}}}\right) + \alpha \hat{\vec{\mathbf{g}}} + \varepsilon$$

A partir de este modelo se puede construir el test C, denominado así porque en él se estima el valor de α condicionado a los valores $\hat{\vec{\boldsymbol{\beta}}}$**.** Este segundo test es recomendado por los autores para los casos en los que la función $f\left(X_i, \vec{\boldsymbol{\beta}}\right)$ es no lineal.

La aplicación del contraste es similar a la del contraste J, con la diferencia de que inicialmente se estiman los dos modelos por separado y posteriormente se contrasta la nulidad del valor α sujeto a los valores obtenidos de $\hat{\vec{\boldsymbol{\beta}}}$.

Por tanto, la validez de un modelo de regresión depende de que la especificación funcional y las variables explicativas utilizadas sean las apropiadas. Una vez realizada esta elección el modelo debe verificar las hipótesis a priori supuestas para el mismo, o por el contrario realizar modificaciones sobre este para éstas se cumplan. Una vez obtenido el modelo de regresión, éste debe tener sentido práctico en el tema que se esté tratando, de forma que los coeficientes del modelo sean coherentes con las variables que se están analizando.

Cuando se halla encontrado el modelo que mejor explica la variable dependiente, éste puede ser utilizado para la estimación de valores Y, a partir de valores inéditos de X en la muestra seleccionada.

Los modelos construidos con este método están limitados por su dependencia a la no existencia de cambios estructurales en los datos analizados, de forma que un suceso concreto en un momento determinado puede modificar las relaciones entre series temporales, o un cambio en la situación espacial de los datos de estudio puede modificar en exceso también estas relaciones.

Una alternativa a este método y que se utiliza con frecuencia desde hace años es la aplicación de una técnica no paramétrica como son las redes neuronales, que tienen el inconveniente principal de la no obtención de un modelo sencillo, pero que a menudo consiguen mejoras significativas en la estimación debido a su capacidad de cálculo en relaciones no lineales

3.Redes neuronales

3.1. Introducción

La inteligencia artificial (IA) se puede definir como la rama de la computación que trata de reproducir razonamientos humanos en sistemas informáticos a través del desarrollo de software y hardware específicos.

El objetivo fundamental de esta disciplina es dotar de inteligencia a la máquina, de forma que, a partir del aprendizaje de determinados problemas, ésta sea capaz de resolver problemas no analizados anteriormente.

Turing (1950), propone una prueba para determinar si un ordenador es inteligente o no. Para ello, una persona, el interrogador, formula un conjunto de preguntas a través de un ordenador para las que esperará una respuesta. Se afirmará que la máquina es inteligente, si el interrogador es incapaz de discernir si las respuestas han sido elaboradas por una persona o una máquina.

La IA es una rama de la computación con un gran potencial, ya que trata de dotar a una máquina de la capacidad para:

- Generalizar un razonamiento a partir de un problema particular dado
- Resolver un problema no resuelto anteriormente.

Ambas capacidades convierten a las máquinas en individuos inteligentes.

Desde sus inicios, esta disciplina ha conseguido importantes avances en multitud de ámbitos que generan a su vez numerosas aplicaciones como el reconocimiento de patrones, el diagnóstico clínico, la robótica, la organización del trabajo o en las finanzas. También la aplicación que nos ocupa en esta tesis, la valoración del precio de un inmueble para unas características determinadas.

Dentro de la Inteligencia artificial se pueden distinguir dos grandes campos: La inteligencia artificial simbólica y la inteligencia artificial subsimbólica.

En la inteligencia artificial simbólica se diseñan sistemas capaces de resolver un problema dado a partir del aprendizaje de la disciplina correspondiente. Es decir, se implementan en la máquina todos los conocimientos propios de una disciplina y se trata de que la máquina los utilice para realizar razonamientos, semejantes a los humanos, que les permitan resolver problemas de aplicación no implementados directamente.

En la inteligencia artificial subsimbólica se parte de un sistema general de resolución de problemas, que es implementado en la máquina, este sistema es adaptado y reconfigurado, por ésta, mediante el aprendizaje, para la resolución de un problema concreto.

Dentro del campo de la inteligencia artificial subsimbólica se engloban las redes neuronales artificiales, objetivo de estudio de este capítulo.

Para definir el concepto de red neuronal artificial, repasemos las definiciones dadas por diversos autores.

Hecht-Nielsen (1989) definió las redes neuronales como un sistema dinámico que tiene la topología de un grafo dirigido y que puede llevar a cabo el procesamiento de la información por medio de su estado de respuesta a una entrada.

Teuvo Kohonen (1988) las definió como redes de elementos simples, generalmente adaptativos, que están interconectadas masivamente en paralelo con una organización

jerárquica, que tratan de interactuar con objetos del mundo actual de forma similar a como lo haría el sistema nervioso biológico.

Caridad y Ceular (2001) las definen como sistemas que tratan de mimetizar la estructura computacional del sistema nervioso humano con el fin de resolver problemas de carácter cognitivo que no son fáciles de programar en modo algorítmico.

En su libro *A comprehensive foundation*, Haykin (1994), ofrece una definición concisa y clara en la que se indica la estructura, funcionamiento y objetivo fundamental de una red neuronal artificial. De esta forma, la define como un procesamiento distribuido masivamente en paralelo que tiene una tendencia natural a almacenar conocimiento empírico y hacerlo disponible para el uso. Recuerda al cerebro en dos aspectos: El conocimiento se adquiere a través de un proceso de aprendizaje, y las conexiones entre elementos (neuronas) se conocen como pesos sinápticos y se usan para almacenar el conocimiento.

De todo ello, entendemos que una red neuronal artificial, es un modelo computacional basado en el funcionamiento del sistema neuronal biológico, en el que la información que se introduce en la red, fluye a través de ésta de forma recurrente, modificándola para obtener un aprendizaje que dé las respuestas deseadas a partir de la información dada.

Las redes neuronales artificiales pueden ser consideradas como técnicas estadísticas no paramétricas. Esta clasificación es considerada por algunos autores como simplista y a menudo genera controversia.

Podemos, al menos afirmar, que un gran número de las aplicaciones importantes de las redes neuronales tienen como objetivo fundamental la aplicación de técnicas estadísticas como son:

- Regresión lineal y no lineal.
- Regresión logística.
- Análisis discriminante.
- Análisis de componentes principales.
- Escalas multidimensionales.

La principal ventaja que presentan las redes neuronales respecto a técnicas paramétricas, es que son aplicables a situaciones en las que las variables de estudio no verifican las hipótesis requeridas en estas últimas. Además, en muchas ocasiones, demuestra obtener mejores resultados que la técnica paramétrica.

Sin embargo, también presenta grandes inconvenientes como la difícil interpretación del funcionamiento de la red, que impide la construcción de modelos estadísticos de estimación. A menudo se hace referencia a este funcionamiento como al de una caja negra.

Las primeras teorías sobre el funcionamiento del cerebro humano datan del siglo IV antes de Cristo, y se atribuyen a Platón y Aristóteles. Éstas tuvieron su continuidad con Descartes en el siglo XVI.

El primer investigador que estudió el funcionamiento del cerebro como modo de funcionamiento computacional, fue Alan Turing, en 1936.

Sin embargo, todos los autores coinciden en situar el inicio de las redes neuronales artificiales en 1943, en el trabajo de Warren McCulloch y Walter Pitts, quienes propusieron un modelo de neurona artificial en la que ésta se activa o no en función de los estímulos o entradas recibidas de otras neuronas. Además, indicaron ya, la capacidad que ésta tiene de aprender. En este trabajo iniciaron el camino a demostrar que la actividad nerviosa y las relaciones neuronales pueden ser descritas a través de la lógica proposicional.

Seis años después, Hebb (1949), desarrolló el procedimiento de aprendizaje de una red neuronal biológica. Fue la base para determinar los métodos para modificar los pesos entre dos neuronas de una red, de manera que la red aprendiese.

En 1951, Marvin Minsky obtuvo resultados prácticos construyendo una máquina denominada SNARC, basada en los modelos de McCulloch y Pitts, que modelizaba el comportamiento de una rata en un laboratorio, y que estaba compuesta por 40 neuronas.

En la conferencia convocada por McCarthy, en 1956, se acuñó por primera vez la denominación inteligencia artificial, y se definieron los objetivos para los próximos años.

Sin embargo, las expectativas creadas en este congreso no fueron satisfechas, motivo por el cual se frenó el avance por el abandono de la disciplina de muchos investigadores.

Unos años más tarde, Rosenblatt (1958), desarrolló el perceptrón simple. Durante años, trabajó en esta red neuronal añadiéndoles capacidad de aprendizaje a los modelos de McCulloch y Pitts, y generando así la primera red neuronal artificial completa. El primer modelo que se desarrolló fue el denominado fotoperceptrón, que era un dispositivo que imitaba al ojo humano, y que permitía, mediante el aprendizaje, clasificar patrones correctamente.

La gran limitación que encontró Rosenblatt en el Perceptrón fue su incapacidad de clasificar clases no linealmente separables, lo que reducía drásticamente su utilidad.

Dos años después, Widrow y Hoff (1960), desarrollaron el modelo Adaline (Adaptative Linear Element), un modelo muy similar al perceptrón de Rosenblatt, pero con aplicaciones completamente distintas. Las diferencias fundamentales radicaban en la forma de utilizar la salida en la regla de aprendizaje y en la gráfica de la función de error, que posee un mínimo absoluto a diferencia del perceptrón, en el que, al tener más de uno, el proceso puede finalizar en un mínimo local. No obstante, este modelo presentaba la misma limitación que el perceptrón: su incapacidad para clasificar clases no linealmente separables.

A partir de ese momento surgieron numerosos estudios sobre redes neuronales como los de Steinbuch, Grossberg, Amari o Anderson. Hasta que, en 1969, Minsky y Papert publicaron su libro *Perceptrons*, en el que consideraban improbable la superación de las limitaciones del perceptrón por la imposibilidad de extender este a perceptrones multinivel. Este estudio marcó un punto de inflexión en el que numerosos investigadores abandonaron este campo.

No obstante, el estudio continuó con el asociador lineal de Anderson y Kohonen, o el neocognitrón de Fukushima.

No fue hasta 1982, cuando se retomaría el interés por la disciplina, impulsado por el desarrollo informático y la implementación de las redes neuronales para aplicaciones en robótica.

El momento del resurgimiento del perceptrón tiene lugar cuatro años más tarde, momento en el que Rumelhart, Hinton y McClelland (1986) desarrollaron el perceptrón multicapa y popularizaron el método de retropropagación y la regla delta generalizada. Esto permitía resolver problemas de clasificación no linealmente separables, a través de la introducción de capas ocultas en la red.

En la Tabla 6 se muestra de forma resumida las principales aportaciones en el campo de las redes neuronales artificiales.

Tabla 6. Principales aportaciones en redes neuronales artificiales

Autor/es	Año	Aportación
Alan Turing	1936	Cerebro como modo de funcionamiento computacional
Warren McCulloch y Walter Pitts	1943	Primer modelo de neurona artificial
Donald Hebb	1949	Primer procedimiento de aprendizaje
Marvin Minsky	1951	Red compuesta por 40 neuronas
John McCarthy	1956	Denominación inteligencia artificial
Frank Rosenblatt	1958	Perceptrón simple
Bernard Widrow y Marcial Hoff	1960	Adaline
Marvin Minsky y Seymour Papert	1969	Determinaron la posibilidad de construir redes multicapa
Rumelhart, Hinton y McClelland	1986	Perceptrón multicapa

Fuente: Elaboración propia

Las aplicaciones de las redes neuronales y los usos prácticos que se le han dado son, desde entonces, innumerables, siendo hoy en día una metodología de resolución de problemas muy extendida a nivel multidisciplinar.

Las redes neuronales han sido ampliamente utilizadas en la valoración de inmuebles desde principios de los años 90, por lo que existen trabajos en diversas zonas geográficas del mundo. En España, los trabajos en este campo tienen sus inicios a principios de siglo. En la Tabla 7, construida a partir de las contenidas en el trabajo de Rey (2014), y actualizada, pueden observarse las principales aportaciones en orden cronológico. Las aplicaciones al mercado inmobiliario español se han resaltado en negrita.

Tabla 7. Principales aplicaciones de las ARN a la valoración de inmuebles.

Autor/es	Año	Aportación	Autor/es	Año	Aportación
Borst	1991	Nueva Inglaterra	Khalafallah	2008	Orlando
Tay y Ho	1992	Singapur	Lam y Yu	2008	Hong Kong
Do y Grudnitski	1992	California	Fernández y otros	2008	**Madrid**
Collins y Evans	1994	Reino Unido	Selim	2009	Turquía
Worzala y otros	1995	Colorado	Peterson y Flanagan	2009	Carolina del Norte
McCluskey y otros	1996	Irlanda	Kauko y otros	2009	Hungría
Shaaf y Erfani	1996	Florida	Shi y otros	2009	China
Rossini	1997	Australia	Wu y otros	2009	Taiwan
Bonissone Cheetham	1997	E.E.U.U.	Kusan y otros	2010	Turquía
Haynes y Tan	1998	Australia	Kamzaoi y Hernández	2011	México
Cechin y otros	2000	Brasil	Kontrimas y Verikas	2011	Lituania
Ceular y Caridad	2000	**Córdoba**	Lin y Mohan	2011	E.E.U.U.
Karakozova	2000	Finlandia	Canavarro	2011	Portugal
Nguyen y Cripps	2001	Tennessee	Zurada y otros	2011	Kentucky
Kauko y otros	2002	Finlandia	Amri y Tularam	2012	Australia
Mohamed	2002	**Cádiz**	McCluskey y otros	2012	Irlanda
Limsombunchai	2004	Nueva Zelanda	Landajo y otros	2012	Oviedo
Fuentes	2004	**Melilla**	Muñoz	2012	**Córdoba**
García	2004	**Albacete**	Fernández y otros	2012	**Valencia**
Gallego	2004	**Madrid**	Mimis y otros	2013	Grecia
Lara	2005	**Jaén**	Azadeh y otros	2014	Irán
Liu, Zhang y Wu	2006	China	Kutasi y Badics	2016	Budapest
Núñez	2007	**Córdoba**			

Fuente: Elaboración propia a partir de Rey (2014)

3.2. Red neuronal biológica

Una red neuronal artificial (RNA) es un sistema computacional que está formado por una estructura que imita las redes neuronales biológicas. El estudio de éstas últimas, analizado por Rosenzweig, Leiman y Breedlove (1998) nos permitirá entender mejor el funcionamiento de una RNA.

El cerebro humano está compuesto por millones de células denominadas neuronas, que están interconectadas entre sí o con otra célula emisora o receptora, mediante unos enlaces químico – eléctricos de unión denominados sinapsis. Una gran cantidad de información procedente de los sentidos llega a las neuronas mediante estas interconexiones, esta información es procesada y comparada con información procesada anteriormente, para dar la respuesta deseada.

La principal característica de este sistema es, precisamente, la capacidad de aprendizaje del mismo.

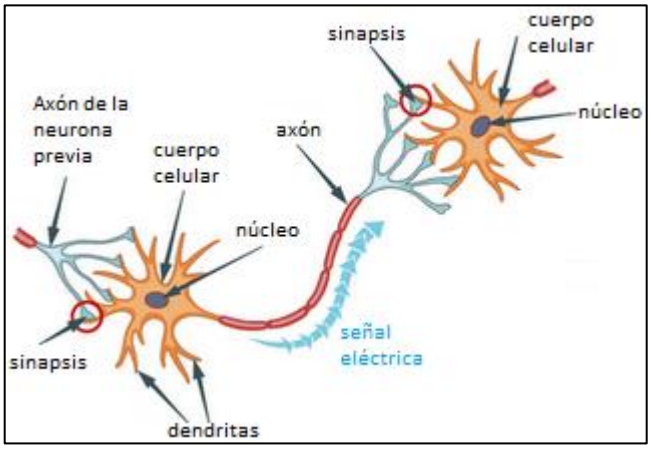


Figura 3. Conexión y comunicación de las neuronas. Fuente: study.com

Como se observa en la *Figura 3*, la neurona consta de un cuerpo celular denominado soma, del que emergen unas ramificaciones denominadas dendritas y una prolongación tubular denominada axón. Ésta recibe un conjunto de señales a través de las dendritas, las procesa mediante impulsos eléctricos en el soma y emite determinada señal a las neuronas a las que está conectada a través de otras dendritas que emergen del axón.

La neurona transmite ondas eléctricas debida a la diferencia de potencial existente entre el interior y el exterior de la célula. Una neurona inactiva tiene polaridad negativa. Cuando la célula se activa, genera un potencial de acción que cambia la polaridad durante unos milisegundos. Si aumenta la diferencia de potencial negativa, se dice que la neurona se ha hiperpolarizado. En caso de disminuir, decimos que la neurona se despolariza. En este último caso, si el estímulo recibido supera un valor umbral, se dispara el denominado potencial de acción, generando un impulso. En caso de no superar este umbral no genera ningún impulso.

En caso de activación de la neurona, una vez generado el potencial de acción, éste alcanza el extremo de la neurona provocando la liberación de unas moléculas denominadas neurotransmisoras, que se fijan en los receptores de otra neurona o de un músculo o glándula generando la siguiente respuesta del sistema.

El sistema nervioso está compuesto por:

- Las neuronas aferentes o neuronas sensitivas que son aquellas que reciben estímulos desde el exterior del circuito neuronal, es decir, desde los receptores u órganos sensoriales.
- Las interneuronas o neuronas de asociación que son las que realizan la interconexión entre otras neuronas del circuito.
- Las neuronas eferentes o neuronas motoras, que son las que transportan los impulsos nerviosos al exterior del circuito. Están conectadas a músculos o glándulas.

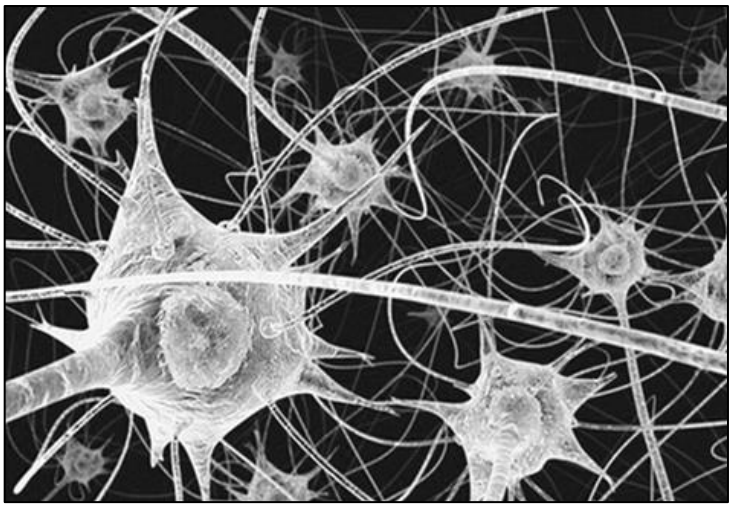


Figura 4. Morfología de las neuronas y sus conexiones. Fuente: pmgbiology.com

Se calcula que en el cerebro humano existe un número de neuronas del orden de 10^{11}, cada una de las cuales está conectada a unas 10.000 neuronas. Esto forma una red de neuronas de enorme tamaño, con una gran capacidad de procesamiento de datos y aprendizaje. Una imagen de esta estructura se puede observar en la Figura 4

3.3. Red neuronal artificial

Como hemos indicado anteriormente, una RNA es un sistema computacional que emula esta estructura biológica. No obstante, el orden de magnitud del número de neuronas que pueden considerarse y de las posibles interconexiones entre ellas, es

sustancialmente más pequeño y viene limitado por la capacidad de procesamiento de los ordenadores. Veamos a continuación la estructura de una red neuronal artificial.

3.3.1. Elementos de una red neuronal artificial

Cada unidad de una RNA, que denominaremos neurona por similitud a la red biológica, se compone de los siguientes elementos:

- Un conjunto de **valores de entrada** $x_1, x_2, \dots, x_n$, que pueden proceder del exterior de la red, en cuyo caso la neurona pertenece a la denominada capa de entrada; o pueden ser un conjunto de respuestas dadas por otras neuronas, con lo que la neurona pertenece a una capa interna o a la capa externa de la red. Cada conjunto diferente e valores de entrada se denomina patrón.
- Un conjunto de valores $\omega_{1j}, \omega_{2j}, \dots, \omega_{nj}$ denominados **pesos sinápticos** que miden la interacción entre cada neurona emisora de información i con $i = 1, \dots, n$, y la neurona receptora, j.
- La **regla de propagación**, que es la función que combina los valores de entrada con los pesos sinápticos para proporcionar un valor real que denominaremos potencial o potencial postsináptico de la neurona. La forma más usada para la regla de propagación es:

$$h_j = \sum_{i=1}^{n} \omega_{ij} \cdot x_i$$

  Es decir, es una función lineal, y mide la suma ponderada de los valores de entrada a la neurona.

- La **función de activación**, que es la función que combina el valor obtenido en la regla de propagación, es decir, el potencial de la neurona, h_j, con un valor que mide el denominado estado de activación actual de la neurona, θ_j^t, para obtener el estado de activación futuro de ésta, θ_j^{t+1}.

  El estado inicial de activación de cada neurona se inicializa al principio del proceso asignándole un valor, en principio, aleatorio. Existen diversas funciones de activación que determinan en buena medida la escala de valor de la salida de la neurona. Las más usadas son:

  - Funciones escalón:

$$f(t) = \begin{cases} 1 & si\ h_j > \theta_j \\ 0 & si\ h_j \leq \theta_j \end{cases} \quad \text{rango } = \{0, 1\}$$

$$f(t) = \begin{cases} 1 & si\ h_j > \theta_j \\ -1 & si\ h_j \leq \theta_j \end{cases} \quad rango = \{-1, 1\}$$

- Funciones sigmoideas:

$$f(t) = \frac{1}{1 + e^{-(h_j - \theta_j)}} \quad rango = [0, 1]$$

$$f(t) = tanh(h_j - \theta_j) \quad rango = [-1, 1]$$

Se utilizará una u otra, en función de la escala deseada en la salida de la neurona. Ésta es común a todas las neuronas.

- La **función de salida**, que es una función que transforma el nuevo estado de activación de la neurona en el valor de salida de ésta, y_j que será el que transmita a las siguientes neuronas, a las que está conectada o al exterior de la red. Ésta suele ser únicamente una función de reescalado, una función escalón o incluso la función identidad.

El esquema de funcionamiento de la neurona artificial se resume en la Figura 5.

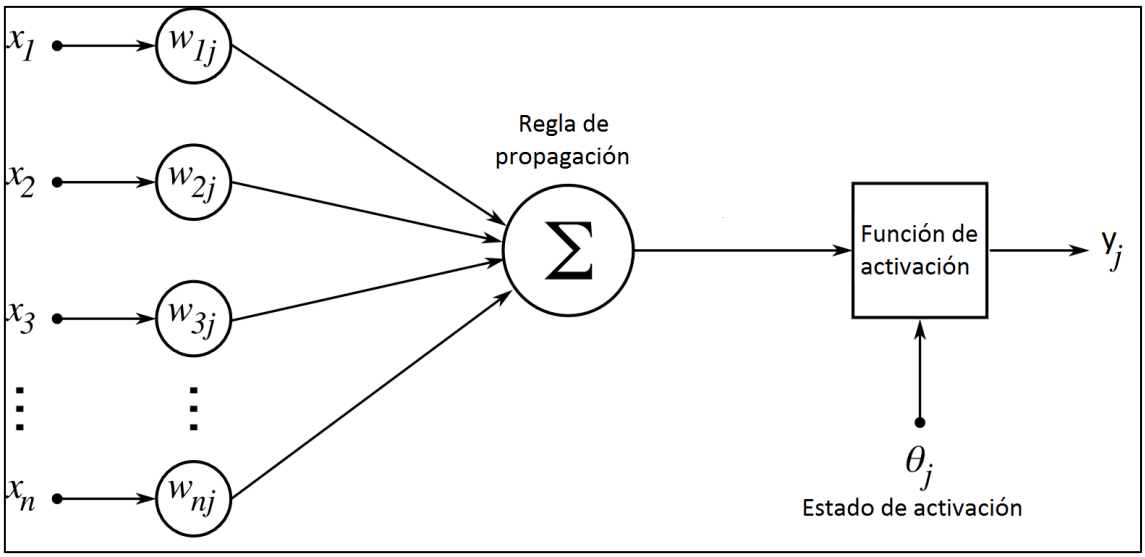


Figura 5. Modelo de una red neuronal artificial. Fuente: Neural Networks Framework

De esta manera, cada neurona de la red recibe un conjunto de valores numéricos de entrada, que pueden proceder del exterior, en cuyo caso se denominan neuronas de la capa

de entrada, o de otras neuronas de la red. A partir de estos valores y de los pesos sinápticos entre las neuronas emisoras y la receptora, se obtiene el potencial de la neurona a través de la regla de propagación. Este potencial, junto con el valor del estado de activación actual de la neurona, produce el nuevo estado de activación, a través de la función de activación. Por último, este valor se transforma en la salida de la neurona a través de la función de salida, que es el valor que emite a las siguientes neuronas de la red, o el valor que devuelve al exterior de la red, en cuyo caso diremos que la neurona pertenece a la capa de salida.

La forma en la que se conectan las neuronas entre sí para formar la red se denomina **arquitectura de red**. Un ejemplo básico de arquitectura es el que se muestra en la Figura 6.

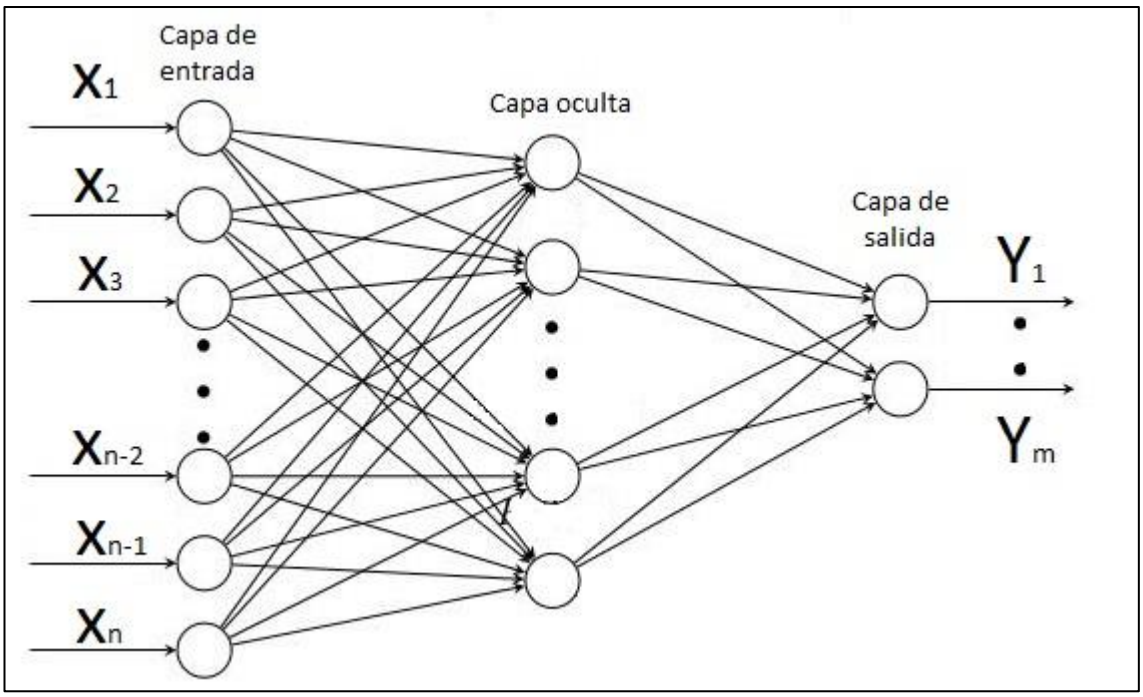


Figura 6. Arquitectura de una red neuronal artificial.

Por tanto, podemos resumir el funcionamiento de una red de la siguiente forma: La red recibe un conjunto de valores $x_1, x_2, \dots, x_n$ que se transmiten a través de las neuronas de la red para devolver un conjunto de valores de salida $y_1, y_2, \dots, y_m$.

Para un conjunto de valores de entrada y unas funciones de propagación, activación y salida fijas, los valores de salida $y_1, y_2, \dots, y_m$, dependen de los pesos sinápticos entre las neuronas, ω_{ij}, y de los estados de activación de éstas. Es decir, si modificamos

convenientemente los pesos sinápticos de una red neuronal, podemos obtener los valores de salida de la red deseados. A este proceso se le denomina proceso de aprendizaje de una red neuronal.

Dados k conjuntos de entradas, el aprendizaje de la red se puede definir como el proceso por el que se modifican los pesos sinápticos entre las neuronas de forma que los k conjuntos de salida correspondientes tengan unas características concretas definidas previamente, y de forma que para conjuntos de entrada distintos a los usados, pero razonablemente parecidos, sea capaz también, de devolver valores de salida con las características deseadas, es decir, que sea capaz de generalizar su comportamiento. La capacidad de aprendizaje de una red neuronal es su característica más importante.

El proceso de aprendizaje de la red se realiza de la siguiente forma: Se introducen todos los ejemplos del conjunto de aprendizaje en la red, y se comprueba si estos verifican un determinado criterio de convergencia. Si no se verifica este criterio, se modifican los pesos sinápticos de la red y vuelven a introducirse los ejemplos en la red. En caso de verificarse, el proceso de aprendizaje finaliza y la red está lista para ser utilizada con nuevos conjuntos de valores de entrada.

El conjunto de aprendizaje es el conjunto formado por todos los conjuntos de valores de entrada utilizados para modificar los pesos sinápticos de la red para que ésta actúe de la forma deseada. Este conjunto debe ser significativo, es decir, debe haber un número de conjuntos de entrada suficientemente grande para que la red sea correctamente entrenada; y debe ser representativo, es decir, deben ser considerados todos los tipos distintos de datos para que la red esté entrenada para las particulares características de cada uno de ellos. Cada conjunto de valores de entrada del conjunto de aprendizaje se denomina ejemplo.

Se pueden distinguir básicamente tres tipos de aprendizaje:

- Aprendizaje supervisado.
- Aprendizaje no supervisado o autoorganizado.
- Aprendizaje híbrido.
- Aprendizaje reforzado.

En el aprendizaje supervisado, para cada conjunto de valores de entrada de la red $x_1, x_2, \dots, x_n$, se tiene información sobre la salida deseada. Esta información es la que permite medir la bondad del resultado $y, y_2, \dots, y_m$ obtenido por la red, y, por tanto, modificar los valores de los pesos sinápticos de manera que los valores de salida de la red verifiquen la información deseada.

Podemos utilizar tres procedimientos de parada para este algoritmo recurrente:

- A partir de una función de error, que mide la discrepancia, fijar una cota por debajo de la cual la discrepancia entre las características de los valores devueltos por la red y las de los valores deseados sea asumible. Una vez alcanzado este valor, el proceso finaliza.
- Fijando a priori un número de ciclos a realizar por la serie, es decir, un número de veces que se va a introducir todo el conjunto de valores de entrada para modificar los pesos sinápticos de la red.
- Detener el algoritmo cuando la variación de los valores de los pesos sinápticos entre dos ciclos consecutivos sea menor que un valor predeterminado.

Un ejemplo de aprendizaje supervisado, que además es el caso que nos ocupa en este estudio, es aquél en el que además de tener el conjunto de valores de la entrada de la red, $x_1, x_2, \dots, x_n$, tenemos el conjunto de valores de salida deseada $s_1, s_2, \dots, s_m$. En este caso, los valores de los pesos se modificarán en cada iteración de manera que la diferencia entre los valores de salida de la red $y_1, y_2, \dots, y_m$ y los valores deseados $s_1, s_2, \dots, s_m$sea mínima. El aprendizaje terminará cuando la diferencia o error esté por debajo de una determinada cota definida con anterioridad, o cuando se alcance el mínimo valor de éste.

En el aprendizaje no supervisado no se dispone de información adicional a la de los valores de entrada de la red. Por este motivo, el ajuste de los pesos sinápticos se realiza únicamente mediante la información de los valores de entrada. La red trata de hallar características de los datos que permitan su clasificación según similitud, para después modificar los pesos de forma que, para dos conjuntos de valores de entrada similares, la red devuelva salidas también parecidas. Es decir, a partir de los resultados dados por la red, ésta modifica sus pesos de manera que, si consideramos los valores de entrada como variables aleatorias, los resultados obtenidos sigan una determinada función de probabilidad o densidad, por lo que la red es capaz de reconocer patrones. Ejemplos de este método de aprendizaje son el aprendizaje de Hebb, en el que el peso entre dos

neuronas aumenta su valor si las dos neuronas han sido activadas en la iteración anterior, y la regla de aprendizaje competitivo.

Cuando se combinan los dos tipos de aprendizaje anteriores, éste se denomina aprendizaje híbrido. Un ejemplo de este tipo de aprendizaje se da cuando no se conocen las respuestas deseadas de la red, pero sí se conoce si la respuesta obtenida coincide o no con ésta.

En el aprendizaje reforzado no se dispone de información sobre las salidas esperadas de la red como en el aprendizaje supervisado. Sin embargo, sí que se dispone de un índice global de rendimiento de la red, es decir, una medida de la bondad de la red. A partir de esta medida, se modificarán los pesos sinápticos con el objetivo de mejorar el valor de este índice.

Por lo tanto, este es un método de aprendizaje en el que se dispone de información que permite la modificación de los pesos por comparación con el resultado deseado. Es por este motivo que algunos autores lo consideran un tipo de método de aprendizaje supervisado.

En Quesada, et al. (1994), los autores recogen en una tabla similar a la que se muestra en la Tabla *8*, los principales algoritmos de aprendizaje utilizados, junto con su clasificación, tiempo de entrenamiento y capacidad de almacenamiento:

Tabla 8.*Propiedades de los diferentes algoritmos de aprendizaje. Fuente: Quesada et al. (1994)*

Algoritmo de aprendizaje	**Tiempo de entrenamiento**	**Método de aprendizaje**	**Capacidad de almacenamiento**
Hebbiano	Rápido	No supervisado	Pobre
Competitivo	Lento	No supervisado	Buena
Min – Max	Rápido	No supervisado	Buena
Corrección de error en dos niveles	Lento	Supervisado	Buena
Corrección de error multinivel	Muy Lento	Supervisado	Muy buena
Reforzado	Súper lento	Supervisado	Buena
Estocástico	Súper Lento	Supervisado	Buena

Fuente: Elaboración propia

Una vez finalizado el proceso de aprendizaje de la red, los pesos sinápticos quedan fijados, así como la estructura de la misma. En este momento la red está preparada para procesar nuevos datos y devolver para éstos una determinada respuesta.

La posición de cada neurona en la red, así como las interconexiones posibles entre neuronas definen la topología de la red neuronal. Si todas las neuronas de una capa están conectadas a todas las de la siguiente capa, se dice que está completamente conectada. Según esta estructura, las redes de neuronas artificiales se pueden clasificar en:

- Redes monocapa con propagación hacia delante. En este caso, la red está compuesta de dos capas, la de entrada en la que las neuronas no realizan operación alguna, sólo transmiten la información a la siguiente capa, y las de salida, encargadas de procesar la información recibida y generar la respuesta. Se dice que la propagación es hacia delante porque las neuronas de la capa de entrada únicamente están conectadas con las de la capa de salida, y nunca consigo misma o con una neurona de su misma capa.
- Redes multicapa con propagación hacia delante. Esta topología de red consta, además de las capas de entrada y de salida, de un número determinado de capas ocultas. Las neuronas de estas capas se denominan neuronas ocultas. El objetivo de añadir capas ocultas a la red, es que esto aumenta considerablemente la capacidad de la red de modelizar cualquier función.
- Redes recurrentes. Este tipo de redes se diferencian de las anteriores en que una neurona determinada puede, no sólo estar conectada con neuronas de la capa siguiente sino con neuronas de capas anteriores, de la misma capa, o incluso consigo misma, de modo que la salida de la neurona se convierte en entrada. Esta topología dinámica la hace especialmente útil para estudiar la dinámica de sistemas no lineales. El inconveniente fundamental es que no siempre es una estructura convergente a un estado estable, por lo que requiere de la verificación de condiciones que aseguren dicha estabilidad.

3.3.2. Ventajas y desventajas de las redes neuronales artificiales

Entre las principales ventajas de las redes neuronales artificiales podemos destacar las siguientes:

- La información en una red es redundante por lo que una RNA puede funcionar correctamente aun cuando parte de ésta no lo haga.
- No necesita de programación de algoritmos ya que utiliza únicamente información sobre los datos de entrada y los datos de salida deseados.
- Las RNA son sistemas dinámicos autoadaptativos, es decir, tiene la capacidad de adaptarse en tiempo real.

- Las RNA realizan transformaciones no lineales de los datos.
- Puede considerarse una técnica no paramétrica en el sentido de que no requiere que las variables de entrada verifiquen una forma determinada.

No obstante, también existe un número considerable de desventajas, entre las que destacamos:

- El tiempo requerido por la red para aprender es desconocido *a priori*.
- El diseño de la red más eficaz no puede determinarse *a priori*, tendrá que ser determinado mediante ensayos de diferentes diseños.
- Normalmente el tiempo de convergencia es lento.
- Una red con subaprendizaje no será capaz de dar buenos resultados, mientras que una red con sobreaprendizaje no será capaz de generalizar a otro conjunto de valores de entrada.
- El conjunto de aprendizaje debe ser suficientemente grande y representativo para poder obtener buenos resultados.
- Existe una gran dificultad en conocer cómo se procesa internamente la información. Aunque es posible obtener la ecuación de red, este proceso es muy laborioso y su complejidad aumenta no linealmente con el número de neuronas y capas de la red.

La red neuronal más utilizada en la actualidad para estimar modelos de regresión es el perceptrón multicapa. Es por ello que a continuación se desarrollará en primer lugar el perceptrón simple para continuar con el perceptrón multicapa.

3.4. El perceptrón simple

La primera RNA, fue diseñada y desarrollada por Rosenblatt (1958), y lo denominó perceptrón. Al contrario que otros investigadores de su época, Rosenblatt se decantó por analizar las relaciones entre las neuronas en términos probabilísticos en lugar de a través de la lógica proposicional.

El perceptrón simple es una red monocapa con propagación hacia delante, es decir, está compuesta por la capa de entrada, formada por n neuronas de entrada, y una capa de salida formada por m neuronas de salida.

Dados un conjunto de N patrones de entrada, consideremos el conjunto de valores de entrada discretos que denotaremos por $x_1^k, x_2^k, \dots, x_n^k$ con $k = 1, \dots, N$, devuelve un

conjunto de valores de salida, que denotaremos $y_1^k, y_2^k, \dots, y_m^k$. La función de activación será de tipo escalón.

Por tanto, la función que relaciona el conjunto de salida con un patrón de entrada dado puede expresarse de la siguiente forma:

$$y_j^k = f\left(h_j^k - \theta_j\right) = f\left(\sum_{i=1}^{n} \omega_{ij} \cdot x_i^k - \theta_j\right) \quad con\ j = 1, \dots, m$$

Al aplicar la función de activación de tipo escalón, queda:

$$y_j^k = \begin{cases} 1 & si \sum_{i=1}^{n} \omega_{ij} \cdot x_i^k > \theta_j \\ 0 & si \sum_{i=1}^{n} \omega_{ij} \cdot x_i^k \leq \theta_j \end{cases}$$

Que es la función original usada por Rosenblatt.

También puede usarse:

$$y_j^k = \begin{cases} 1 & si \sum_{i=1}^{n} \omega_{ij} \cdot x_i^k > \theta_j \\ -1 & si \sum_{i=1}^{n} \omega_{ij} \cdot x_i^k \leq \theta_j \end{cases}$$

Como puede observarse, la respuesta de la red es dicotómica, por lo que esta red neuronal se utiliza como clasificador o discriminante lineal de un conjunto de N puntos $\left(x_1^k, x_2^k, \dots, x_n^k\right)\ con\ k = 1, \dots, N$ en $\mathbb{R}^n$, ya que como puede observarse:

$$\sum_{i=1}^{n} \omega_{ij} \cdot x_i^k = \theta_j$$

Es la ecuación de un hiperplano en $\mathbb{R}^n$.

En este hecho radica la gran limitación del perceptrón simple. Es un clasificador con capacidad de aprendizaje que sólo puede ser utilizado para discriminar entre dos clases

linealmente separables. Hecho que fue ampliamente criticado por Minsky y Papert (1969) y que provocó un gran freno a la investigación en redes neuronales durante años. A continuación, se muestra un ejemplo en el plano, dado en este mismo trabajo, que utilizaron los autores para explicar la incapacidad del perceptrón en estos casos. Figura 7.

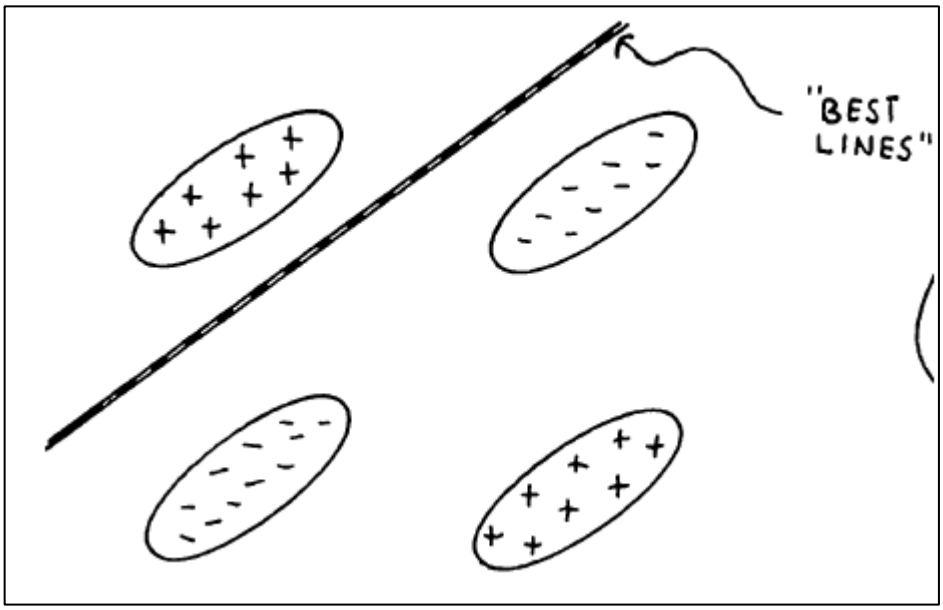


Figura 7. Perceptrón simple. Discriminador de clases linealmente separables. Fuente: Minsky y Papert (1969)

Como hemos visto, para un patrón concreto, el valor devuelto por el perceptrón simple, en el caso en el que sólo tengamos una neurona de salida, es 1 o -1 (0 si escogemos la primera función de activación). Si el este valor coincide con el valor esperado, los pesos permanecerán constantes, pero si el valor no es el esperado, se actualizará el valor de éstos.

Si tanto los valores de entrada como los de salida sólo toman los valores 1 y -1, el cambio en el valor de los pesos es el siguiente:

$$\Delta\omega_{ij} = \eta \cdot \left(y_j^k - s_j^k\right) \cdot x_i^k$$

Donde η se denomina ritmo de aprendizaje. Éste tomará un valor comprendido entre 0 y 1. Es importante la elección del valor de η debido a que valores muy pequeños pueden hacer que la convergencia de la red sea muy lenta ya que produce que el cambio en los pesos sea también reducido, mientras que valores muy elevados pueden hacer oscilar a la

red. De esta forma los valores de los pesos en el instante $t + 1$ en función de los pesos de la red en el instante t serán:

$$\omega_{ij}(t+1) = \omega_{ij}(t) + \sum_{k=1}^{N} \Delta\omega_{ij}^{k}(t)$$

Este método de aprendizaje, que como hemos visto anteriormente, es un método supervisado denominado algoritmo de aprendizaje Hebbiano, comentado anteriormente.

Rosenblatt (1958) demostró que, independientemente de los valores iniciales dados a los pesos, si los conjuntos son linealmente separables, el algoritmo converge en un número finito de iteraciones.

3.5. El perceptrón multicapa

Minsky and Papert (1969), no sólo se limitaron a hacer una crítica demoledora de la utilidad del perceptrón por su incapacidad de separar espacios no lineales. También expusieron, que esta limitación podría ser superada combinando varios perceptrones simples.

No obstante, esta propuesta quedó en el aire ya que carecían de un método de actualización de pesos aplicable a redes construidas de esta forma, y para las que la regla de aprendizaje del perceptrón simple no era aplicable.

Fue diecisiete años más tarde, en 1986, cuando James L. McClelland y David E. Rumelhart, junto al grupo de investigación PDP, publicaron su trabajo, en el que usando como base la arquitectura propuesta por Minsky y Papert, desarrollaron un método de aprendizaje de la red, propagando el error de ésta hacia atrás en la red, por lo que se denominó algoritmo de propagación hacia atrás o *backpropagation*, que fue el precursor de la regla delta generalizada.

En la Figura 8, extraída de Minsky and Papert (1969) se muestra la arquitectura del perceptrón multicapa, bajo la cual los autores expresan que “debería existir un algoritmo de aprendizaje que optimice los pesos a partir del error de estimación”, pero los autores no lo habían investigado.

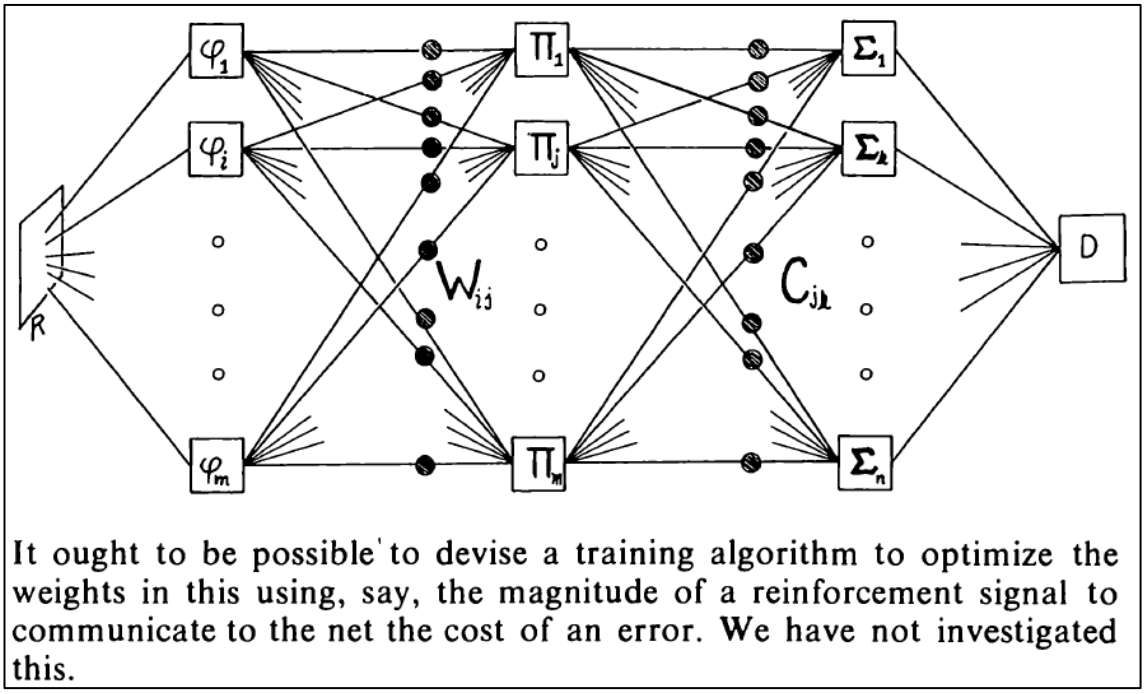


Figura 8.Arquitectura del perceptrón multicapa. Fuente: Minsky y Papert (1969)

Toda función continua en el espacio real n – dimensional puede aproximarse a través de un perceptrón multicapa. Diferentes autores han demostrado que, en general, una única capa oculta o a lo sumo dos, es suficiente.

La estructura de una red de tipo perceptrón multicapa está compuesta por:

- La capa de entrada, que está formada por n neuronas. Estas neuronas reciben directamente la información de los patrones de entrada, y la transmiten a las neuronas de la siguiente capa.
- Un conjunto de capas intermedias denominadas capas ocultas. Las neuronas de estas capas reciben la información de las capas de entrada, la transforman y la transmiten a las neuronas de la siguiente capa oculta o a las de la capa de salida.
- La capa de salida, que está formada por m neuronas. Estas neuronas reciben la información de la última capa oculta, la transforman generando la salida de la red.

Ésta es una red multicapa de propagación hacia delante, por lo que una neurona cualquiera de la red sólo recibe información de neuronas de capas anteriores, y la emite a neuronas de la siguiente capa. Por tanto, no tendrá conexiones laterales, es decir, con neuronas de su misma capa, ni con neuronas de capas anteriores. Es, además, una red con aprendizaje supervisado, por lo que se conocen los valores esperados para la salida de la red, y por tanto el error cometido.

A partir de ahora, supondremos, por simplificar la notación, una red con una única capa oculta. La generalización a un mayor número de capas ocultas es inmediata.

Sean $x_1^k, x_2^k, \dots, x_n^k$ con $k = 1, \dots, N$, el conjunto de N patrones de entrada, que vamos a utilizar para entrenar la red, $s_1^k, s_2^k, \dots, s_m^k$ y el conjunto de valores esperados de salida de la red, para cada patrón de entrada $x_1^k, x_2^k, \dots, x_n^k$.

Denotamos por ϑ_{ij} con $i = 1, \dots, n$ y $j = 1, \dots, h$ a los pesos sinápticos entre la neurona i de la capa de entrada y la neurona j de la capa oculta; y por ω_{jt} con $j = 1, \dots, h$, y $t = 1, \dots, m$ a los pesos entre la neurona j de la capa oculta, y la neurona t de la capa de salida.

Como vimos en el perceptrón simple, la salida dada por una neurona j de la capa oculta a un patrón de entrada dado $x_1^k, x_2^k, \dots, x_n^k$ es:

$$z_j^k = f_j\left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1}\right) con\ j = 1, \dots, h$$

Donde f_j y θ_{j1} son respectivamente, la función de activación y el potencial de activación de la neurona j de la capa oculta.

De la misma forma, para una entrada $z_1^k, z_2^k, \dots, z_h^k$ desde la capa oculta a la de salida, la red devuelve:

$$y_t^k = f_t\left(\sum_{j=1}^{h} \omega_{jt} \cdot z_j^k - \theta_{t2}\right) con\ t = 1, \dots, m$$

Donde f_t y θ_{t2} son respectivamente, la función de activación y el potencial de activación de la neurona t de la capa de salida.

Por tanto, para el patrón de entrada $x_1^k, x_2^k, \dots, x_n^k$, la salida de la red es:

$$y_t^k = f_t\left(\sum_{j=1}^{h} \omega_{jt} \cdot f_j\left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1}\right) - \theta_{t2}\right) con\ t = 1, \dots, m$$

Una vez obtenida la salida de la red, ésta se compara con los valores $s_1^k, s_2^k, \dots, s_m^k$ deseados, para calcular la medida del error cometido. La función de error más utilizada en esta red neuronal es la dada por el método LMS (Least Mean Square) propuesto por Widrow y Hoff (1960). Para el patrón k, el error se define como:

$$\varepsilon^k = \frac{1}{2} \cdot \sum_{t=1}^{m} \left(s_t^k - y_t^k\right)^2$$

La modificación de los pesos de la red puede realizarse de dos formas:

- Aprendizaje en serie: Los pesos son actualizados tras la presentación de cada patrón de entrada. En este caso, la introducción de patrones en la red debe realizarse de forma aleatoria. Tiene, en general, un mayor coste computacional.
- Aprendizaje por lotes: Se actualizan los pesos una vez se han presentado todos los patrones de aprendizaje. Se calcula el error total cometido por la red como la suma de los errores de cada patrón. Es el método más habitual.

El error total es, por tanto:

$$\varepsilon = \frac{1}{2} \cdot \sum_{k=1}^{N} \sum_{t=1}^{m} \left(s_t^k - y_t^k\right)^2$$

Nótese que, en última instancia, ε depende de los pesos de la red, por lo que el vector opuesto a su gradiente respecto a estos, da la dirección de máximo decrecimiento de la función. De esta forma, el cambio de los pesos se calculará como la derivada parcial del error respecto a cada peso cambiada de signo, multiplicada por la tasa de aprendizaje.

$$\Delta\vartheta_{ij} = -\eta \cdot \frac{\partial \varepsilon}{\partial \vartheta_{ij}}$$

$$\Delta\omega_{jt} = -\eta \cdot \frac{\partial \varepsilon}{\partial \omega_{jt}}$$

Así, de la función del error:

$$\varepsilon = \frac{1}{2} \cdot \sum_{k=1}^{N} \sum_{t=1}^{m} \left(s_t^k - y_t^k\right)^2 = \frac{1}{2} \cdot \sum_{k=1}^{N} \sum_{t=1}^{m} \left(s_t^k - f_t \left(\sum_{i=1}^{n} \omega_{jt} \cdot z_j^k - \theta_{t2} \right) \right)^2$$

Se obtiene que los pesos entre las neuronas de la capa oculta y las de la capa de salida, cambian a partir de la expresión:

$$\Delta\omega_{jt} = -\eta \cdot \frac{\partial \varepsilon}{\partial \omega_{jt}} = \eta \cdot \sum_{k=1}^{N} \left(s_t^k - y_t^k\right) \cdot f_j' \left(\sum_{i=1}^{n} \omega_{jt} \cdot z_j^k - \theta_{t2} \right) \cdot z_j^k$$

Si denotamos por

$$\delta_t^k = \left(s_t^k - y_t^k\right) \cdot f_j' \left(\sum_{i=1}^{n} \omega_{jt} \cdot z_j^k - \theta_{t2} \right)$$

La expresión del incremento queda:

$$\Delta\omega_{jt} = \eta \cdot \sum_{k=1}^{N} \delta_t^k \cdot z_j^k$$

De la misma forma, para los pesos entre las neuronas de la capa de entrada y las de la capa oculta el incremento se calcula a partir de la expresión de error:

$$\varepsilon = \frac{1}{2} \cdot \sum_{k=1}^{N} \sum_{t=1}^{m} \left(s_t^k - y_t^k\right)^2$$

$$= \frac{1}{2} \cdot \sum_{k=1}^{N} \sum_{t=1}^{m} \left(s_t^k - f_t \left(\sum_{i=1}^{n} \omega_{jt} \cdot f_j \left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1} \right) - \theta_{t2} \right) \right)^2$$

Por lo que la expresión del incremento queda:

$$\Delta\vartheta_{ij} = -\eta \cdot \frac{\partial \varepsilon}{\partial \vartheta_{ij}} = \eta \cdot \sum_{k=1}^{N} f_j \left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1} \right) \cdot \sum_{j=1}^{m} \delta_t^k \cdot \omega_{jt} \cdot x_i^k$$

Usando la notación:

$$\delta_j^k = f_j \left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1} \right) \cdot \sum_{t=1}^{m} \delta_t^k \cdot \omega_{jt}$$

Por lo que:

$$\Delta\vartheta_{ij} = \eta \cdot \sum_{k=1}^{N} \delta_j^k \cdot x_i^k$$

Por tanto, el valor δ_j^k de cualquier neurona de la capa oculta depende de los δ_t^k de las neuronas de la capa de salida. Por tanto, se puede decir que el error se propaga a lo largo de la red hacia atrás. Por este motivo, este algoritmo se conoce como *backpropagation* o de retropropagación.

De esta forma, los pesos de la red se actualizarán tras la presentación de todos los patrones mediante las expresiones:

$$\vartheta_{ij}(iter.+1) = \vartheta_{ij}(iter.) + \Delta\vartheta_{ij}$$

$$\omega_{jt}(iter.+1) = \omega_{jt}(iter.) + \Delta\omega_{jt}$$

La actualización del valor del potencial de activación de las neuronas se hace de la misma forma, teniendo en cuenta que este es un peso más de la neurona que siempre recibe una entrada igual a -1.

$$\Delta\theta_{t2} = -\eta \cdot \frac{\partial\varepsilon}{\partial\theta_{t2}} = -\eta \cdot \sum_{k=1}^{N} \left(s_t^k - y_t^k\right) \cdot f'\left(\sum_{i=1}^{n} \omega_{jt} \cdot z_j^k - \theta_{t2}\right)$$

$$\Delta\vartheta_{ij} = -\eta \cdot \frac{\partial\varepsilon}{\partial\vartheta_{ij}} = -\eta \cdot \sum_{k=1}^{N} f_j\left(\sum_{i=1}^{n} \vartheta_{ij} \cdot x_i^k - \theta_{j1}\right) \cdot \sum_{j=1}^{m} \delta_t^k \cdot \omega_{jt}$$

El método de actualización usado aquí, tiene el inconveniente de que puede converger a un mínimo de error relativo de la función y no al mínimo absoluto, por lo que una vez entrenada una red, ésta deberá ser testada para comprobar su eficacia.

McClelland, Rumelhart et al (1986), propusieron añadir un factor al aprendizaje, denominado momento de aprendizaje, que permitiese filtrar las oscilaciones y acelerar significativamente la convergencia. De esta forma, el algoritmo de aprendizaje quedaría de la siguiente forma:

$$\Delta\vartheta_{ij}(iter.+1) = \eta \cdot \sum_{k=1}^{N} \delta_j^k \cdot x_i^k + \alpha \cdot \Delta\vartheta_{ij}(iter.)$$

$$\Delta\omega_{jt}(iter.+1) = \eta \cdot \sum_{k=1}^{N} \delta_t^k \cdot z_j^k + \alpha \cdot \Delta\omega_{jt}(iter.)$$

Como hemos indicado anteriormente, el valor de la tasa de aprendizaje, η, es un valor comprendido entre 0 y 1. Valores muy próximos a 1 harían del algoritmo un proceso lento, debido a que la modificación de los pesos en cada iteración es muy limitada. Sin embargo, un valor cercano a 1 podría llevar a importantes oscilaciones en el método. Por este motivo, los valores que suelen utilizarse son los comprendidos entre 0.05 y 0.5.

En McClelland, Rumelhart et al (1986), se propone, para el momento de aprendizaje, valores cercanos a 1.

3.6. Aplicación del perceptrón multicapa en problemas de predicción

El perceptrón multicapa puede ser utilizado para la resolución de problemas de clasificación y de predicción. En esta tesis se utilizará, como se ha mencionado anteriormente, esta última aplicación.

Los pasos a seguir para la resolución de un problema de predicción a través de la construcción de una red neuronal de tipo perceptrón multicapa, son los siguientes:

- Fase de identificación.
- Entrenamiento de la red.
- Fase de validación.
- Interpretación de pesos y estimación
- Reentrenamiento de la red, en caso de que los resultados no sean los óptimos, modificando los parámetros de la arquitectura de la red.

Veamos en detalle cada uno de ellos.

3.6.1. Fase de identificación

Esta es la fase en la que se define el modelo que se desea estimar, indicando la variable o variables dependientes, así como el conjunto de variables explicativas que se utilizará para explicarlas.

Como en los modelos de regresión, es importante seleccionar un buen conjunto de variables explicativas que verifique dos condiciones: Todas las variables deben ser relevantes para estimar la variable dependiente, y que a su vez no exista una relación fuerte entre ellas, ya que esto puede provocar sobreajuste del modelo.

En nuestro caso, la variable dependiente será el precio del inmueble, y el conjunto de variables explicativas estará compuesto por las variables indicadoras de las características intrínsecas del inmueble como localización, tamaño o calidades.

Para analizar la relevancia individual de cada variable se construirá la red con todas las variables explicativas y se medirá el error cometido por ésta. A continuación, se eliminará una de las variables explicativas y se comparará el error cometido con el de la primera red. Este proceso se realiza para todas las variables explicativas. La variable que al ser eliminada genere un menor incremento en el error de la red, siendo este no significativo, será considerada irrelevante.

Una vez seleccionadas las variables que se van a incluir en el modelo, se normalizan sus valores al rango $[0, 1]$ si se va a utilizar como función de activación la función logística, y el intervalo $[-1, 1]$ si se va a usar la tangente hiperbólica. Los valores de las variables dicotómicas se asignarán a los extremos del intervalo, y las cualitativas no dicotómicas se introducirán en la red mediante variables artificiales con tantas neuronas como categorías menos 1.

En esta fase, además, se dividirá la muestra en tres submuestras: muestra de entrenamiento a partir de la cual se actualizarán los pesos de la red, muestra de validación que se utilizará para evitar el sobreaprendizaje de la red, ya que una vez estimados los pesos, la eficacia se comprobará para otros patrones distintos a los de entrenamiento; y muestra de test, con la que se evaluará la capacidad predictiva de la red.

3.6.2. Entrenamiento de la red

Para llevar a cabo el entrenamiento de la red, se deben definir lo siguientes parámetros:

- Arquitectura de la red. El número de neuronas de la capa de entrada viene determinado por el número de variables explicativas. El número de neuronas de la capa de salida, viene determinado por el número de variables dependientes, en nuestro caso, una. Por tanto, sólo queda definir el número de capas ocultas de la red, y el número de neuronas en cada una de estas capas. Como se ha comentado anteriormente, la mayoría de los problemas pueden resolverse con una única capa oculta. Además, un número elevado de neuronas en la capa oculta podría provocar un sobreajuste en la red, lo que implicaría que la red perdería la capacidad de generalización. No existe un criterio unificado para determinar el número óptimo de estas, por lo que se calculará probando diferentes arquitecturas.
- Inicialización de pesos. La elección de los pesos iniciales de la red se puede realizar de forma aleatoria, pero es conveniente utilizar valores en el intervalo $[-0.5, 0.5]$ y de forma que la regla de propagación devuelva un valor próximo a 0. De esta forma, la función de activación tomará valores alejados de los extremos, donde la pendiente es reducida, y por tanto el aprendizaje lento.
- Elección de la función de activación de las neuronas. La ventaja fundamental entre la estimación mediante modelos de regresión y perceptrón es la capacidad de este último de explicar relaciones no lineales. Esto será posible siempre que en las neuronas de la capa oculta se definan funciones de activación no lineales. Es por ello que se suelen utilizar en éstas, funciones sigmoideas como la función logística o la tangente hiperbólica. En problemas de predicción suelen utilizarse la función identidad en las neuronas de la capa de salida.
- Asignación de los valores de la tasa y el momento de aprendizaje. Se ha comentado anteriormente la importancia de estos factores en la velocidad de aprendizaje de la red. La tasa de aprendizaje debe definirse como un valor entre 0.05 y 0.5, mientras que se deben tomar valores próximos a 1 para el momento de aprendizaje.

3.6.3. Fase de validación

En esta fase se determina la bondad de la red entrenada en la fase anterior. Para esta fase es utilizada la submuestra de test. La medida más utilizada es la del error cuadrático medio.

Es recomendable entrenar diferentes arquitecturas de red, y comparar los resultados obtenidos para así validar el modelo que menor error cuadrático medio devuelva, siempre que éste sea menor o igual al deseable.

3.6.4. Interpretación de pesos y estimación

Una vez seleccionada la red óptima, ya puede utilizarse para la estimación de valores de la variable dependiente. En nuestro caso, la red debe estimar con precisión el precio de un inmueble a partir de características no estudiadas anteriormente.

El análisis del peso que cada variable tiene en la red sobre la variable o variables dependientes ha sido una de las principales críticas que ha sufrido esta metodología, debido a la complejidad de éstas. No obstante, numerosos estudios han ido dirigidos a determinar el peso que cada variable explicativa tiene sobre la variable dependiente.

Se dice que una red está identificada, si no presenta neuronas irrelevantes, tanto en la capa de entrada como en las capas ocultas. White (1989) demuestra que, bajo estas condiciones, los pesos de la red seguirán una distribución Normal multivariante, lo que permite la construcción de un intervalo de confianza para los pesos de la red óptima, y, por tanto, de la importancia relativa de cada neurona de la red en la estimación de la variable dependiente. Si además la función de error es la función de log-verosimilitud negativa.

Montaño, Palmer et al. (2002), realizan una revisión de los diferentes métodos que se han propuesto, y presentan un nuevo método, que a continuación se resumen.

Existen dos métodos, el análisis basado en la magnitud de los pesos, y el análisis de sensibilidad.

En los métodos basados en la magnitud de los pesos, para redes con una capa oculta, se trata de medir la importancia que cada variable explicativa tiene para explicar la variable dependiente a partir de los pesos sinápticos de la red, una vez finalizado su aprendizaje. Para cada neurona i de la capa de entrada se calcula la suma de los productos de los pesos entre esta neurona y cada una de las neuronas de la capa oculta, y el peso entre la neurona de la capa oculta y la neurona de la capa de salida. El índice más usado

según Montaño, Palmer et al. (2002) es el propuesto por Garson (1991) que con la notación vista anteriormente, para una única neurona en la capa de salida y considerando los valores absolutos de los pesos, es:

$$Q_i = \frac{\sum_{j=1}^{h} \frac{\vartheta_{ij}}{\sum_{i=1}^{n} \vartheta_{ij}} \cdot \omega_{j1}}{\sum_{i=1}^{n} \sum_{j=1}^{h} \frac{\vartheta_{ij}}{\sum_{i=1}^{n} \vartheta_{ij}} \cdot \omega_{j1}}$$

Tal y como se ha calculado el índice es obvio que la suma de los n índices correspondientes a cada una de las variables de entrada suman 1, y, por tanto, éste puede tomarse como porcentaje de la variable dependiente que es capaz de explicar determinada variable de entrada.

El análisis de sensibilidad consiste en realizar pequeñas modificaciones en el valor de una de las variables explicativas dejando el resto de variables constantes para analizar la variación producida en el resultado dado por la red. Es lo que se puede denominar un análisis marginal.

En el trabajo mencionado anteriormente, Montaño, Palmer et al. (2002) revisan los siguientes métodos:

- Análisis de sensibilidad basado en el error. Se utiliza medida de relevancia la variación de la raíz del error cuadrático medio, de forma que una vez entrenada la red, se realizan pequeñas variaciones, a lo largo de su rango de definición, de los valores de una de las variables explicativas, dejando el resto invariante. Se entrena la variable y se calcula nuevamente el valor de la raíz del error cuadrático medio. La variable para la que se produzca una mayor variación de este índice será la más relevante. El inconveniente de este método es que sólo permite medir la relevancia de variables continuas.
- Análisis de sensibilidad basado en la salida. Estos métodos evalúan la relevancia de cada variable explicativa a partir de la variación producida en el valor de la variable dependiente cuando se realiza pequeñas variaciones en el valor de ésta, a partir de, por ejemplo, los valores medios del resto de variables explicativas. Dentro de este análisis de sensibilidad los autores destacan el método de la matriz de sensibilidad Jacobiana y el método de sensibilidad numérico, propuesto por ellos. El primero de ellos calcula para cada variable de entrada, x_i, su valor de sensibilidad como:

$$S_i = \frac{\partial y}{\partial x_i}$$

El inconveniente de este método es que para su aplicación es necesario asumir la normalidad de las variables. Por otra parte, en el método de sensibilidad numérico (SNA), se ordenan en sentido creciente de x_i los patrones de entrenamiento. Se subdivide en grupos de tamaños similares y se calculan los valores medios de x_i y de y de cada subgrupo. El índice de sensibilidad entre dos grupos consecutivos g_r y g_{r+1} se define como:

$$NSA_i(g_r) = \frac{\bar{y}(g_{r+1}) - \bar{y}(g_r)}{\bar{x}_i(g_{r+1}) - \bar{x}_i(g_r)}$$

Se define el índice de sensibilidad de x_i como el valor esperado de los índices de cada grupo.

La selección de la arquitectura óptima de la red no es un tema trivial, y ésta debe realizarse mediante la comprobación de diferentes arquitecturas. Como se ha mencionado previamente, en la mayoría de los casos, una única capa oculta suele ser suficiente, y un elevado número de neuronas en la capa oculta puede generar sobreaprendizaje de la red, que conllevaría una menor capacidad de generalización de ésta. Por este motivo, se recomienda comenzar con redes neuronales con una capa oculta y un número reducido de neuronas en la capa oculta que puede incrementarse en busca de un mejor resultado.

La capacidad predictiva de la red hallada se calcula a través de la muestra de validación. No obstante, Murata, Yoshizawa et al. (1994) proponen la generalización, del criterio de información de Akaike para modelos de regresión, a redes neuronales, denominado criterio de información de la red: NIC, con el que mide la capacidad de ajuste de la red a la muestra dada.

Este criterio, al igual que el definido en modelos de regresión, está compuesto por dos términos. El primero de ellos mide el valor esperado de la discrepancia entre el modelo obtenido y el modelo teórico que a partir de los datos dados obtiene el valor de la variable dependiente. El segundo, penaliza la mayor complejidad de la red.

4.Valores atípicos

4.1. Introducción

En este capítulo vamos a tratar con detenimiento la existencia de observaciones en los datos de estudio que puedan modificar sustancialmente la capacidad predictiva y la bondad de ajuste del modelo que se desea construir. Realizaremos una revisión de los métodos univariantes y multivariantes, ampliamente utilizados por diversos paquetes estadísticos como SPSS, para la detección de valores atípicos y profundizaremos, para el segundo tipo, en un método basado en el cálculo del criterio de información de un modelo de regresión, tratado en algunos trabajos previos como el de Kornacki, Kyureghyan, y Ignaciuk (2012) o el de Karagrigoriou, Mattheou, y Vonta (2011).

En el mercado inmobiliario, puede tener tanta importancia, o más, el análisis de observaciones que puedan considerarse anómalas, como la propia construcción del modelo de regresión, ya que el primero nos permitirá detectar inmuebles que pueden considerarse significativamente diferentes, en algún sentido, a aquellos de su entorno, y por ende con características que hacen relevante su distinción respecto a los de sus mismas características, como puede ser un precio significativamente más bajo o más alto de lo que indican sus características.

La gran heterogeneidad de los sujetos de estudio de este mercado, reduce en gran medida la capacidad predictiva de los modelos de predicción de precios, por lo que es de suma importancia la detección de observaciones atípicas, para su posible interpretación, y tras la que el experto decida su inclusión o no en el estudio de un modelo que tiene como objetivo explicar los inmuebles más característicos de la zona.

El objetivo principal es detectar inmuebles que por sus características perjudiquen a la estimación del resto en su conjunto, evitando introducir este sesgo en el modelo. Por tanto, si hallamos un inmueble con un precio muy elevado, trataremos de justificar esto a través de otras variables de estudio como la localización, el tamaño, el tipo de inmueble, o los servicios adicionales que pueda tener como cochera o piscina. Tras realizar un minucioso análisis y si no se han hallado causas que justifiquen este valor, el investigador podrá decidir eliminarlo del estudio, siempre con precaución para evitar el sesgo en el sentido contrario que esto pudiera introducir.

Identificar inmuebles sobresalientes es además de gran valor para detectar aquellos con características propias diferenciadoras que pueden ser utilizadas, por ejemplo, publicitariamente, para mejorar su probabilidad de venta; o para hacer constar al propietario del inmueble la existencia de deficiencias en su oferta que puedan impedir o dificultar la venta de este.

Este es el motivo por el que el método basado en la variación del criterio de información es tan importante en el análisis del mercado inmobiliario, ya que, como observaremos detenidamente, éste detecta únicamente las observaciones que se desvían significativamente de la nube de puntos en el sentido de que su distancia al hiperplano de regresión es significante superior al resto, y no las que se alejan de los puntos en una dirección paralela al hiperplano de regresión.

Sin embargo, métodos como, por ejemplo, el basado en el cálculo de los valores palanca o *leverage*, detecta únicamente las observaciones alejadas de la nube de puntos, aunque su distancia al hiperplano de regresión sea mínima.

4.2. Métodos clásicos de detección de valores atípicos

Definimos valor atípico como aquel que no se ha generado de la misma forma que el resto de los valores de la muestra. Una observación de este tipo puede darse en una única variable, dependiente o independiente, en varias, o incluso ser considerada atípica en su conjunto debido a fuertes discrepancias entre los valores de las variables en su conjunto.

Por tanto, una observación atípica es aquella que puede tomar un valor muy superior (o inferior) al resto, en alguna de las variables analizadas, o como conjunto, pero no necesariamente tiene que modificar el modelo de regresión de forma significativa. Por ejemplo, un inmueble con un elevado precio, pero que esté situado en la mejor zona de la ciudad y con una gran superficie, puede considerarse que tiene un tamaño y precio atípicos, pero en su conjunto puede no influir en exceso en la estimación del modelo de regresión, por no desplazarse de la nube de puntos.

Una observación con valores atípicos en sus variables puede además ser influyente en la estimación del modelo de regresión, si distorsiona de manera significativa la estimación de los coeficientes del modelo o el vector de predicción de éste. A continuación, analizaremos algunos de los métodos utilizados para detectar valores atípicos e identificar si éstos son influyentes o no en la estimación de un modelo de regresión.

El método basado en el cálculo del rango intercuartílico sirve para detectar valores que se diferencian significativamente del resto de valores de una variable concreta, ya sea la variable dependiente o una de las independientes. En éste, consideramos valores atípicos a aquellos menores que $Q_1 - k(Q_3 - Q_1)$ o mayores que $Q_3 + k(Q_3 - Q_1)$, donde Q_1 y Q_3 son los cuartiles y k es un valor al que usualmente se le asigna el valor 1.5 para valores atípicos considerados leves, es decir, que pueden estar generados por la propia estructura de la variable aunque destacable por su valor, y 3 para graves, aquellos cuya existencia es difícil de justificar por la propia naturaleza de la variable.

Por otro lado, el método de la tipificación es un método alternativo al anterior. En él, se consideran valores atípicos aquellos que se desvían de la media de la variable más de un número de veces la desviación típica. Este valor suele tomarse como 2 para valores atípicos leves y 3 para graves.

Ambos métodos pueden considerarse métodos de estudio univariantes ya que sólo tienen en cuenta la presencia de valores fuera de lo normal en alguna de las variables sin profundizar en la influencia que sobre el modelo tienen. El primero de los métodos, al estar basado en cuantiles, la presencia de valores atípicos en la muestra, no tendrá influencia relevante en el cálculo de la frontera que delimita la existencia de éstos, sin embargo, la media y la desviación típica sí son estadísticos muy influenciables, lo que puede generar grandes diferencias en el cálculo de la frontera al excluir estos valores.

Analizaremos, a continuación, la existencia de valores atípicos en la estimación de un modelo de regresión lineal. Todos estos métodos fueron recopilados por Chatterjee y Hadi (1986) y pueden encontrarse de forma habitual en los paquetes estadísticos más utilizados. A continuación, realizaremos una breve revisión de los mismos.

4.2.1. Método del cálculo de valores leverage

Dadas las variables explicativas $X_1, X_2, \dots, X_k$, y un conjunto de observaciones $(x_{1i}, x_{2i}, \dots, x_{ki})\ con\ i = 1, \dots, n$, decimos que la observación $x_j = (x_{1j}, x_{2j}, \dots, x_{kj})$ tiene efecto palanca (*leverage*) sobre el modelo o simplemente, es un punto palanca si éste tiene capacidad de atraer a la ecuación de regresión, y por tanto, su presencia en el conjunto de datos puede llegar a ser influyente.

El efecto palanca de una observación puede medirse como la discrepancia o distancia entre la observación $x_j = (x_{1j}, x_{2j}, \dots, x_{kj})$ y el resto de observaciones consideradas. Para el cálculo de ésta, se utiliza la matriz:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Que es la que mide, a su vez, la matriz de covarianzas de $\hat{Y}$ y por tanto de $\hat{\varepsilon}$, ya que:

$$\boldsymbol{\Sigma}_{\hat{\mathbf{Y}}} = \sigma^2\mathbf{H} \qquad \text{y} \qquad \boldsymbol{\Sigma}_{\hat{\boldsymbol{\varepsilon}}} = \sigma^2(\mathbf{I} - \mathbf{H})$$

De esta forma, definimos

$$\mathrm{h_j} = (\mathrm{x_{1j}}, \mathrm{x_{2j}}, \dots, \mathrm{x_{kj}})(\mathbf{X}'\mathbf{X})^{-1}(\mathrm{x_{1j}}, \mathrm{x_{2j}}, \dots, \mathrm{x_{kj}})'$$

Que puede expresarse, para cada observación $\vec{\mathbf{x}}_{\mathbf{j}} = (x_{1j}, x_{2j}, \dots, x_{kj})$, como una medida del grado de apalancamiento de esta observación sobre el valor estimado para dicha observación. El efecto mínimo de palanca de una observación es igual a $1/n$ y por el contrario su valor máximo es 1. Por ello, valores elevados de efecto palanca, indican la existencia, a priori, de influencia en el modelo de regresión.

Generalmente, según Hoaglin y Welsch (1978), se considera efecto palanca elevado de una observación si su valor supera:

$$\frac{2(k+1)}{n}$$

Siendo k el número de variables explicativas del modelo.

También pueden detectarse valores elevados de efecto palanca con el método de tipificación sobre los valores de apalancamiento de las observaciones, es decir, aquellos que se desvíen de la media, un número de veces determinado, la desviación típica de los valores de palanca.

No obstante, un valor con un elevado efecto palanca, no necesariamente es influyente en la estimación del modelo de regresión, ya que para ello deberá ser causante de cambios significativos en la estimación de los coeficientes de éste.

Decimos que una observación es influyente si al eliminarla del modelo produce en él cambios significativos. Es decir, diremos que lo es si se verifican tres condiciones:

- Produce cambios significativos en la estimación de los coeficientes del modelo.
- Modifica el vector de predicciones.
- El error al estimar el valor de esta observación es muy pequeño cuando ésta forma parte del conjunto de datos original, y muy grande cuando es una observación que no se ha incluido para la estimación del modelo.

4.2.2. Análisis de los residuos. Método de Bonferroni

Para un modelo de regresión dado, podemos analizar la existencia de valores atípicos en éste a través del vector de residuos estimados: $\hat{\varepsilon}$. Si Dada la observación x_j, si ésta es

atípica, el valor de $\widehat{\varepsilon_j}$, obtenido al estimar el valor de la variable dependiente para la observación x_j, será muy elevado respecto al resto de valores de los residuos. Para realizar esta comparación es necesario estandarizar los residuos.

Lo habitual, sobre todo si es la observación tiene un elevado efecto palanca, es *estudentizar* los residuos para una óptima comparación. Esto se realiza mediante la transformación:

$$\hat{\varepsilon}_{S\,j} = \frac{\widehat{\varepsilon_j}}{\hat{\sigma}\sqrt{1-h_j}}$$

Que sigue una distribución *t – Student* con $n-k-2$ grados de libertad bajo los supuestos del modelo de regresión lineal. Un valor elevado será signo de una mala estimación del valor de la variable dependiente para la observación en estudio. Suele tomarse el valor 2 como el límite de desviación asumida, o contrastar su existencia mediante el correspondiente test a partir de la distribución conocida.

No obstante, como el objetivo es determinar la observación con mayor desviación, no es correcto aplicar un test t simple, por lo que la probabilidad límite de este contraste debe ajustarse, ya que la realización de múltiples contrastes sin el ajuste de la probabilidad límite podría llevar a errores en la interpretación de los resultados. Esto es lo que se propone en el contraste ajustado de determinación de valores atípicos de *Bonferroni*, en el que se ajusta la probabilidad límite del contraste univariante por otra en la que se considera el número de contrastes (observaciones) realizado.

Un método alternativo a la aplicación del test con el ajuste propuesto por *Bonferroni*, es aplicar el método de tipificación a los residuos estandarizados, de forma que se consideren residuos atípicos y sus observaciones asociadas, a aquellos cuyo valor se desvíe de la media un número determinado de veces la desviación típica.

Vamos a analizar a continuación la influencia de una observación en la estimación tanto de los coeficientes del modelo de regresión como en las predicciones de la variable dependiente. Para ello, denotamos por $\vec{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k\right)^t$ al vector de las estimaciones de los coeficientes del modelo, y por $\vec{\hat{\beta}}(-\mathrm{i}) = \left(\hat{\beta}(-i)_1, \hat{\beta}(-i)_2, \dots, \hat{\beta}(-i)_k\right)^t$ al vector

de coeficientes estimados cuando eliminamos la observación i-ésima del conjunto de datos. De la misma forma, denotamos la predicción dada por el modelo para el conjunto completo de datos como $\vec{\hat{Y}}$, y al vector de predicción cuando es eliminada la observación i-ésima como $\vec{\hat{Y}}(-\mathrm{i})$.

La medida de influencia de una observación nos la dará la diferencia entre $\vec{\hat{\beta}}$ y $\vec{\hat{\beta}}(-\mathrm{i})$, por un lado, y la de $\vec{\hat{Y}}$ y $\vec{\hat{Y}}(-\mathrm{i})$ por el otro.

4.2.3. Análisis de los valores Dfbetas del modelo

Belsley, Kuh, y Welsch (1980) definen el valor *dfbeta* tipificado del coeficiente β_j para una observación i, como:

$$dfbeta_{ji} = \frac{\hat{\beta}_j - \hat{\beta}(-i)_j}{\hat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$$

Mide la distancia estandarizada entre el valor estimado de $\hat{\beta}_j$ y el de $\hat{\beta}(-i)_j$. Es decir, mide el cambio producido en cada coeficiente del modelo de regresión al eliminar la observación. Valores elevados de *dfbeta* indica que la observación i es influyente.

Pueden ser, por tanto, consideradas observaciones influyentes, aquellas para las que se obtienen valores de *dfbeta* superiores a

$$dfbeta = \left|\frac{2}{\sqrt{n}}\right|$$

4.2.4. Cálculo de la distancia de Cook

El inconveniente de los coeficientes *dfbeta*, radica en la gran cantidad de estimaciones a realizar, ya que habrá una para cada observación eliminada y cada coeficiente de regresión. Es decir, para una muestra de n observaciones y k variables explicativas tendrán que estimarse un total de $n \cdot (k+1)$ coeficientes *dfbeta*. Es por ello que en el

caso de modelos con un gran número de variables explicativas, es recomendable utilizar la distancia de Cook, propuesta por Cook (1977), quién definió una medida global de la distancia entre la estimación de los coeficientes a partir de todas las observaciones y la resultante de eliminar la observación i-ésima. Esta es:

$$CD_i = \frac{\left(\vec{\hat{\beta}} - \vec{\hat{\beta}}(-\mathrm{i})\right)' \mathbf{X}'\mathbf{X}\left(\vec{\hat{\beta}} - \vec{\hat{\beta}}(-\mathrm{i})\right)}{(\mathrm{k}+1)\hat{\sigma}^2}$$

O equivalentemente:

$$CD_i = \frac{\hat{\varepsilon}_{S_i}^2 h_i}{(k+1)(1-h_i)}$$

Para determinar si una distancia es significativamente elevada, podemos comparar la distancia de Cook con el cuantil $\mathcal{F}(k+1, n-k-1)_\alpha$ con un nivel de significación α. Suelen considerarse valores influyentes a aquellos que superan el valor $\mathcal{F}(k+1, n-k-1)_{0.05}$.

Además, como

$$\mathbf{X}\left(\vec{\hat{\beta}} - \vec{\hat{\beta}}(-\mathrm{i})\right) = \vec{\hat{\mathrm{Y}}} - \vec{\hat{\mathrm{Y}}}(-\mathrm{i})$$

Podemos reescribir la distancia de Cook de la siguiente forma

$$CD_i = \frac{\left(\vec{\hat{\mathrm{Y}}} - \vec{\hat{\mathrm{Y}}}(-\mathrm{i})\right)' \left(\vec{\hat{\mathrm{Y}}} - \vec{\hat{\mathrm{Y}}}(-\mathrm{i})\right)}{(k+1)\hat{\sigma}^2} = \frac{1}{(k+1)\hat{\sigma}^2}\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}(-i)_j\right)^2$$

Por lo que la distancia de Cook es, a su vez, una medida de la distancia entre los vectores de valores estimados de Y al incluir o eliminar la observación i-ésima. Nos da un diagnóstico de eliminación de observaciones influyentes.

4.2.5. Cálculo de los valores Dffits del modelo

Otra forma de medir la influencia de una observación sobre un modelo, consiste en calcular la diferencia entre la estimación dada por este modelo cuando ésta forma parte del conjunto de datos, con la que se obtiene al eliminar la observación correspondiente. Para cuantificar esto, Belsley et al. (1980) definieron los coeficientes dffits.

$$Dffits_i = \frac{\hat{Y}_i - \hat{Y}(-i)_i}{\hat{\sigma}\sqrt{h_i}}$$

Mide el número de desviaciones típicas que se desvía la estimación al eliminar la observación i-ésima. Una observación puede considerarse que es un posible valor con influencia en el modelo si dffits verifica:

$$|dffits| > 2\sqrt{\frac{k+1}{n}}$$

4.2.6. Cálculo de los COVratios

Por último, podemos analizar la existencia de valores influyentes desde el punto de vista de la precisión global del modelo. De esta forma definimos el covratio, como el cociente entre el determinante de la matriz de covarianzas de los coeficientes del modelo al eliminar la observación i-ésima y el determinante de ésta sin eliminarla. Así:

$$COVratio_i = \frac{\left|\hat{\sigma}(-i)^2\left(\boldsymbol{X}(-\boldsymbol{i})'\boldsymbol{X}(-\boldsymbol{i})\right)^{-1}\right|}{|\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|}$$

Operando en esta expresión, obtenemos la siguiente, equivalente:

$$COVratio_i = \frac{(\hat{\sigma}(-i)^2)^{k+1}}{(\hat{\sigma}^2)^{k+1}}\frac{1}{1-h_i}$$

Pueden considerarse observaciones influyentes a aquellas para las que el valor absoluto de COVratio cumple:

$$|COVratio| > 1 + 3\frac{k+1}{n}$$

4.2.7. Estudio gráfico de la presencia de valores atípicos

Es recomendable, en el estudio de los valores atípicos combinar la aplicación de algunas de las técnicas vistas hasta este momento con representaciones gráficas que puedan aportar información adicional.

Una representación gráfica sencilla y de gran utilidad son los gráficos QQ, descritos anteriormente, ya que la desviación de determinadas observaciones de la bisectriz del primer cuadrante puede ser indicativo claro de la existencia de valores atípicos.

Para medir el efecto palanca en modelos de regresión simple, basta con superponer, en un gráfico de dispersión, las rectas de regresión, una de ellas, incluyendo todas las observaciones disponibles y la otra eliminando la observación que se sospecha pueda tener efecto palanca.

No obstante, una vez calculados los valores palanca, o las distancias de Cook, o los valores de las probabilidades límite corregidas de los residuos estandarizados, de todas las observaciones, pueden, también, representarse estos en un diagrama de barras en el que puedan detectarse variaciones en las alturas de éstas por un diferente valor.

4.3. Detección de observaciones influyentes basado en el valor de AIC

Tras revisar los métodos diagnósticos de detección de observaciones influyentes en el modelo de regresión lineal, se analiza en profundidad un método basado en el criterio de información de Akaike (AIC), y la variación del valor dado para éste al eliminar del conjunto de datos de análisis, una observación influyente. Uno de los primeros estudios en los que se planteó el análisis de AIC para la detección de valores atípicos fue el de Kitagawa (1979), y a este, le han seguido otros, como por ejemplo, Kitagawa y Akaike (1982), y los mencionados anteriormente. Karagrigoriou et al. (2011) y Kornacki et al. (2012).

En este capítulo analizaremos en profundidad las diferencias en el cambio del valor de AIC en ausencia o presencia de una o más observaciones influyentes.

El inconveniente del análisis estadístico de los criterios de información es la imposibilidad de asociarlos a una distribución conocida que permitiría la construcción de pruebas de hipótesis de detección. Es por ello que se propone el análisis de simulación, mediante el método Montecarlo, de los estadísticos a utilizar.

En primer lugar, se demostrará que el cambio producido en el valor de AIC cuando se elimina una observación en el modelo es significativamente mayor cuando la observación es influyente, y a continuación, se comparará este método con los analizados anteriormente.

4.3.1. El criterio de información de Akaike

El criterio de información de Akaike (1977) nos da una estimación relativa de la información perdida al estimar el conjunto de valores dado a través del modelo escogido en lugar de utilizar la función teórica. Su fórmula, para un modelo con k variables explicativas es:

$$AIC = -2lnL + 2(k+1)$$

Donde lnL es el logaritmo de la función de verosimilitud, supuesta la distribución conocida.

Suponiendo que la distribución de las perturbaciones del modelo verifica que:

$$\vec{\varepsilon} \sim N\left(\vec{0}_n, \sigma_\varepsilon^2 \mathbf{I_n}\right)$$

Y a partir de la función de densidad conjunta de la distribución Normal multivariante, para un conjunto independiente $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, se tiene que:

$$lnL = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2$$

Por lo que:

$$-2lnL = nln(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2$$

Partiendo de que el estimador de máxima verosimilitud de la varianza residual es la varianza muestral, es decir:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Se tiene que:

$$-2lnL = nln(2\pi\hat{\sigma}_\varepsilon^2) + n$$

Por lo que:

$$AIC = nln(2\pi) + nln(\hat{\sigma}_\varepsilon^2) + n + 2(k+1)$$

Que, operando, es equivalente a:

$$AIC = n[ln(2\pi) + 1] + nln(\hat{\sigma}_\varepsilon^2) + 2(k+1)$$

Esta es, también, la fórmula usada por la función AIC del lenguaje R para el cálculo del valor del criterio de información de Akaike.

Dado un conjunto de datos que se genera a través de un proceso desconocido, si este conjunto de datos es generado a través de otro proceso conocido, según Akaike (1977), el valor de AIC podría definirse como una medida de la cantidad de información perdida por usar el segundo proceso, en lugar del primero.

Esta divergencia entre el modelo generador desconocido y el modelo que se utiliza para realizar la estimación, se conoce como divergencia de Kullback – Leibler.

Por tanto, el valor de AIC, es una medida de discrepancia entre el ajuste de datos realizado y el ajuste teórico, por lo que a menor valor de AIC, menor será la discrepancia o pérdida de información, y, por tanto, mejor será el ajuste realizado por el modelo.

Debido a que AIC es, por así decirlo, una medida de discrepancia entre los datos y el modelo utilizado, tradicionalmente, no se recomienda la comparación de valores de AIC en modelo que sirven para estimar conjuntos de datos diferentes.

No obstante, el método de detección de valores atípicos aquí desarrollado, propone, precisamente, la introducción de pequeñas variaciones en los datos, para el posterior análisis de su influencia sobre el valor de AIC.

4.3.2. Variación en el valor de AIC

El objetivo es detectar valores atípicos que, por su naturaleza, generen una elevada influencia en la estimación de los coeficientes del modelo por estar fuera de la curva de tendencia del resto de valores observados.

En la Figura 9 se muestra el gráfico de dispersión de un conjunto de datos pertenecientes a dos variables x e y. Como puede observarse, los valores extremos de la nube de puntos podrían considerarse valores atípicos debido a que toman valores extremos en ambas variables. Sin embargo, hay un valor para el que x toma el valor 10, que puede considerarse, además, atípico por desviarse significativamente de la relación entre vista entre x e y. La presencia de esta observación es influyente en la estimación del modelo de regresión debido a que está situado claramente fuera de la nube de puntos, a una distancia significativamente superior al resto de la recta de regresión.

Por tanto, el método, tiene como objetivo principal, detectar este tipo de observaciones, para decidir acerca de su eliminación, en el caso de que puedan considerarse que el proceso generador de éstas es distinto al del resto de observaciones estudiadas, por lo que su inclusión generaría un empeoramiento en la capacidad de ajuste del modelo.

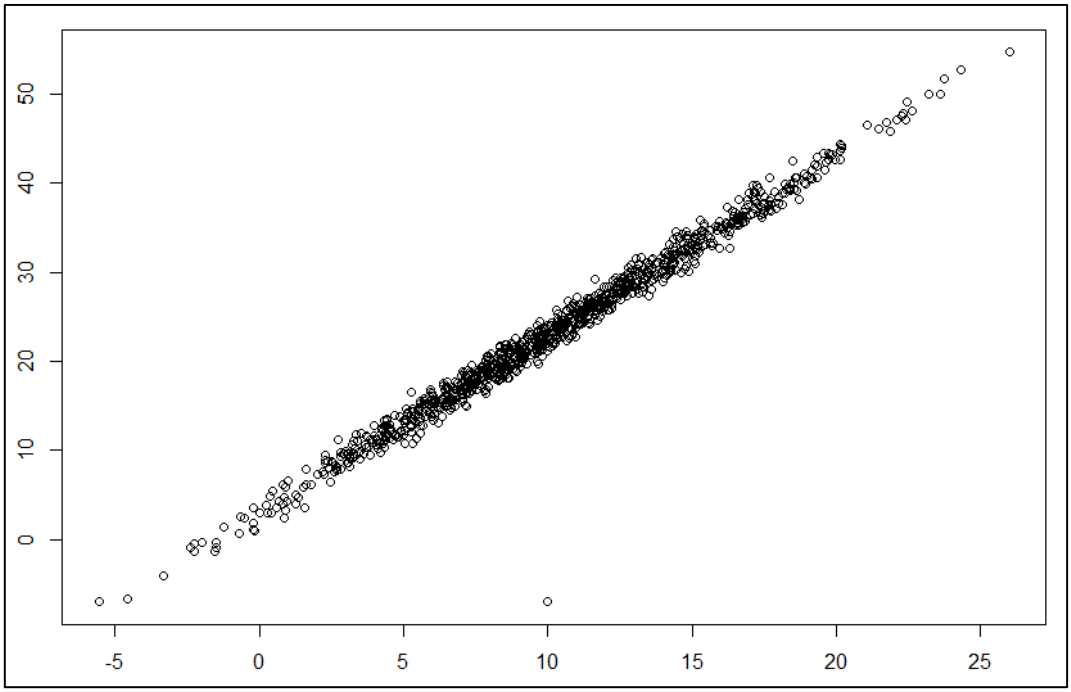


Figura 9. Gráfico de dispersión. Valores atípicos influyentes y no influyentes.

Las observaciones de la Figura 9 que se sitúan en los extremos de la nube, aunque relevantes por ser valores extremos de las variables, no influirán significativamente en la estimación del modelo, y, por tanto, su exclusión no es tan transcendente en la estimación del modelo como las primeras.

En la metodología de precios hedónicos, en el que el objetivo es determinar el valor de un bien a partir del valor de las características que lo conforman, las observaciones influyentes tienen una especial importancia ya que pueden desvirtuar el valor relativo de cada característica del bien estudiado.

Dado un conjunto de variables explicativas $X_1, X_2, \dots, X_k$, una variable Y que se desea estimar a partir de las primeras, y un conjunto de observaciones $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})\ con\ i = 1, \dots, n$, estimamos el modelo de regresión:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

Y el criterio de información de Akaike de dicho modelo: AIC.

Consideraremos, ahora, la observación $j - ésima\ (y_j, x_{1j}, x_{2j}, \dots, x_{kj})$.

Si n es suficientemente grande y esta observación no es influyente, la eliminación de ésta del conjunto de datos, dará lugar a un modelo sin cambios significativos, y, a un valor del criterio de información, $AIC(-j)$, similar a AIC.

El análisis del cambio en el valor del criterio de información al eliminar una observación, debe hacerse con precaución, debido a que el valor de éste depende en gran medida de la escala en la que se miden las variables, el número de éstas e incluso del tamaño muestral.

Para el caso en el que el j-ésimo dato sea eliminado, la muestra estará formada por $n-1$ datos y el valor de AIC será:

$$AIC(-j) = (n-1)[ln(2\pi)+1] + (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-j)\right) + 2(k+1)$$

Denotando por $\hat{\sigma}_\varepsilon^2(-j)$ a la estimación de la varianza residual del modelo obtenido al estimar el conjunto formado por n-1 observaciones.

Si calculamos la diferencia entre ambos valores:

$$AIC - AIC(-j) = ln(2\pi) + 1 + nln(\hat{\sigma}_\varepsilon^2) - (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-j)\right)$$

Que puede reescribirse de la siguiente forma:

$$AIC - AIC(-j) = ln(2\pi) + 1 + ln\left(\hat{\sigma}_\varepsilon^2(-j)\right) + nln(RATIO)$$

Donde:

$$RATIO = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2(-j)}$$

Afirmamos, y analizaremos detenidamente a continuación, que si n es suficientemente grande y la observación i no es influyente, se tiene que:

$$\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2(-j)} \cong 1$$

Por lo que la diferencia quedaría de la siguiente forma:

$$D_i = AIC - AIC(-j) \cong ln(2\pi) + 1 + ln\left(\hat{\sigma}_\varepsilon^2(-j)\right)$$

Como se ha indicado anteriormente, el valor del criterio de información es dependiente de la muestra, por lo que hemos optado por analizar una medida relativa del cambio en el valor de criterio: la proporción de cambio, que es

$$Proporción\ de\ cambio = \frac{AIC - AIC(-j)}{AIC}$$

$$Proporción\ de\ cambio = \frac{ln(2\pi) + 1 + ln\left(\hat{\sigma}_\varepsilon^2(-j)\right)}{nln(2\pi) + nln(\hat{\sigma}_\varepsilon^2) + n + 2(k+1)}$$

Que puede descomponerse en:

$$\frac{ln(2\pi) + 1}{n(ln(2\pi) + 1) + nln(\hat{\sigma}_\varepsilon^2) + 2(k+1)} + \frac{ln\left(\hat{\sigma}_\varepsilon^2(-j)\right)}{nln(\hat{\sigma}_\varepsilon^2) + nln(2\pi) + n + 2(k+1)}$$

Que, evidentemente, tiende a 0 cuando n es suficientemente grande. Esta será la medida que tomaremos para analizar el cambio entre los valores de AIC de las dos muestras.

Cabe destacar, que podríamos haber utilizado el estimador insesgado para la estimación de la varianza residual, es decir:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \bar{s}_\varepsilon^2$$

No obstante, la fórmula de la diferencia hubiera sido la misma, salvo que las estimaciones de las varianzas residuales serían las insesgadas. No obstante, estas diferencias son pequeñas cuando la muestra es suficientemente grande, y por tanto no influirá la utilización de uno u otro estimador de la varianza, por lo que continuaremos con la estimación de la varianza dada por el método de máxima verosimilitud.

Veamos también los cambios que introducirían los distintos criterios de información, en la formulación propuesta. Como se ha visto anteriormente, el criterio de información de Akaike puede escribirse de la siguiente forma:

$$AIC = n[ln(2\pi) + 1] + nln(\hat{\sigma}_\varepsilon^2) + 2(k+1)$$

En él, se penaliza la inclusión de un mayor número de variables explicativas en el modelo de regresión, mediante un coeficiente de penalización igual a 2. Otro criterio, propuesto por (Schwarz, 1978), considera un valor de penalización dependiente del tamaño muestral, e igual a $ln(n)$, denominado criterio de información bayesiano.

$$BIC = n[ln(2\pi) + 1] + nln(\hat{\sigma}_\varepsilon^2) + ln(n)(k+1)$$

Este criterio, por tanto, penaliza el número de parámetros a estimar en el modelo de regresión en mayor medida que el criterio propuesto por (Akaike, 1977), para muestras de tamaño superior a 7, por lo que en la práctica lo consideraremos mayor. Para la muestra que se obtiene al eliminar la observación j, este valor queda:

$$BIC(-j) = (n-1)[ln(2\pi) + 1] + (n-1)ln\big(\hat{\sigma}_\varepsilon^2(-j)\big) + ln(n-1)(k+1)$$

Por lo que su diferencia queda:

$$BIC - BIC(-j) = ln(2\pi) + 1 + ln\big(\hat{\sigma}_\varepsilon^2(-j)\big) + nln(RATIO) + (\boldsymbol{k+1})\boldsymbol{ln}\left(\frac{\boldsymbol{n}}{\boldsymbol{n-1}}\right)$$

El último término de la expresión difiere del resultado calculado para el criterio de información de Akaike. No obstante, cuando n es suficientemente grande, éste tiende a 0, por lo que, ambos términos coinciden. En este caso, la proporción de cambio es:

$$Proporción\ de\ cambio \cong \frac{ln(2\pi) + 1 + ln\big(\hat{\sigma}_\varepsilon^2(-j)\big)}{nln(2\pi) + nln(\hat{\sigma}_\varepsilon^2) + n + ln(n)(k+1)}$$

Que es menor, que la proporción de cambio hallada para el criterio de información de Akaike.

Generalizando ahora el valor de coeficiente de penalización del número de parámetros estimados, y suponiendo que éste toma un valor igual a P, el valor del criterio de información queda:

$$AIC_P = n[ln(2\pi) + 1] + nln(\hat{\sigma}_\varepsilon^2) + P(k+1)$$

Y la diferencia:

$$AIC_P - AIC_P(-j) = ln(2\pi) + 1 + ln\left(\hat{\sigma}_\varepsilon^2(-j)\right) + nln(RATIO) + (k+1)\left(P - P(-j)\right)$$

Que, para valores de penalización invariantes para la eliminación de una observación en la muestra, es igual al calculado para el criterio de información de Akaike. El valor de la proporción de cambio es:

$$Proporción\ de\ cambio \cong \frac{ln(2\pi) + 1 + ln\left(\hat{\sigma}_\varepsilon^2(-j)\right) + (k+1)\left(P - P(-j)\right)}{nln(2\pi) + nln(\hat{\sigma}_\varepsilon^2) + n + P(k+1)}$$

4.3.3. Análisis del Ratio y la proporción del cambio en AIC.

En esta sección analizaremos la veracidad de dos afirmaciones realizadas el apartado anterior, necesarias para la justificación del funcionamiento del método propuesto para la detección de observaciones influyentes. Las hipótesis a comprobar son:

- Cuando n es suficientemente grande:

$$\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2(-j)} \cong 1$$

- La proporción de cambio experimentado por AIC al eliminar una observación, es decir,

$$Proporción\ de\ cambio = \frac{AIC - AIC(-j)}{AIC}$$

  toma valores suficientemente pequeños cuando el tamaño muestral es suficientemente grande, y significativamente menores si la observación eliminada no es influyente

Para ello se ha diseñado un experimento Montecarlo, método propuesto por Hendry (1984) y ampliamente utilizado por distintos autores como Dios (1998). En éste, se trata de demostrar la veracidad de afirmaciones de difícil demostración teórica, como es el caso que nos ocupa, mediante la simulación de variables aleatorias.

Para las simulaciones, se ha optado por la construcción de un modelo de regresión lineal simple, en muestras de tamaño $n = 1000, n = 500\ y\ n = 100$, de forma que

analicemos como la variación del tamaño muestral afecta en el cálculo de ambos términos.

El modelo de regresión considerado es, por tanto:

$$Y = a + bX + \varepsilon$$

La variación en el término independiente del modelo genera un aumento en el cambio de escala entre X e Y, lo que, a su vez genera un aumento en la ratio cuando existe una observación influyente. La media de la variable X, genera, por el contrario, un incremento en la escala global. Además, el cambio en el coeficiente b del modelo produce un cambio en la varianza de la variable dependiente, ya que por la relación lineal definida se tiene que:

$$s_Y^2 = b^2 s_X^2 + s_\varepsilon^2$$

Luego, como:

$$\mathrm{R}^2 = 1 - \frac{s_\varepsilon^2}{s_Y^2} = 1 - \frac{s_\varepsilon^2}{b^2 s_X^2 + s_\varepsilon^2}$$

$$\frac{s_\varepsilon^2}{b^2 s_X^2 + s_\varepsilon^2} = 1 - \mathrm{R}^2$$

$$s_\varepsilon^2 = (1 - \mathrm{R}^2)(b^2 s_X^2 + s_\varepsilon^2)$$

Despejando queda:

$$s_X^2 = \frac{\mathrm{R}^2}{b^2(1 - \mathrm{R}^2)} s_\varepsilon^2$$

Usando esta relación, analizaremos los cambios producidos para distintos valores de a, b, μ_X, R^2 y s_ε^2. Con éstos, el valor de s_Y^2 quedará determinado.

4.3.3.1. Variación en el caso de observaciones no influyentes

En primer lugar, comprobaremos que cuando se introduce una observación no influyente, el valor de la ratio es un valor próximo a 1 y la proporción de cambio es un valor cercano a 0. En diferentes simulaciones realizadas, la posición respecto al eje x de la observación, no ha revelado cambios significativos si ésta se encuentra a la misma distancia de la recta, por lo que se tomará el punto medio del rango de X.

El diseño realizado para simular esta situación es el siguiente:

- Se genera aleatoriamente, con tamaño $n-1$, la variable ε de forma que siga una distribución Normal con media 0 y varianza residual la seleccionada.
- A partir de la ecuación $s_X^2 = \frac{R^2}{100(1-R^2)} s_\varepsilon^2$, con media μ_X, se genera la variable X como una distribución Normal, nuevamente de tamaño $n-1$.
- Se genera la variable Y a partir del modelo lineal elegido: $Y = a + bX + \varepsilon$. Se estima el modelo de regresión, y se calcula el criterio de información de Akaike, $AIC(-i)$, su varianza residual, $s_\varepsilon^2(-i)$, y la media y la desviación típica de las distancias de los puntos a las rectas.
- A continuación, se crea la observación n – ésima, que será una observación no influyente. Esta observación será de la forma (x, y), con:

  $$x = \frac{x_{max} + x_{min}}{2}$$

  e y el valor estimado por la recta de regresión creado en el punto 3 más una perturbación aleatoria Normal con la misma distribución que la variable ε creada en el punto 1. Es decir:

  $$y = \hat{y} + N(0, s_\varepsilon^2)$$

- Esta observación se añade a la muestra creada en los puntos 2 y 3, y con ésta, se genera el nuevo modelo para el que se calculará la varianza residual, $s_{\varepsilon N}^2$, el valor del criterio de información, AIC_N, y la ratio entre $s_{\varepsilon_N}^2$ y $s_\varepsilon^2(-i)$. También, la proporción de cambio entre este modelo y el calculado en el punto 3.
- Este procedimiento se realizará un total de 1000 veces por cada configuración probada. La ejecución devolverá, por tanto, dos vectores formados por estas mismas componentes. El primero de ellos contendrá los valores de las ratios, el segundo, las proporciones de cambio. Se analizarán dichos vectores, calculando para ellos la media, la desviación típica, y se representará gráficamente su comportamiento.

En primer lugar, analizaremos el efecto, que el cambio en el tamaño muestral, tiene sobre la ratio de las varianzas residuales de los dos modelos y sobre la proporción de

cambio del valor del criterio de información de Akaike. Para ello, se fijará el resto de coeficientes y se realizarán diferentes simulaciones para valores crecientes de n. En las siguientes simulaciones se analizará el efecto del cambio de uno de los parámetros fijados, con el objetivo de analizar su influencia. Se comprobará que el resto de factores no cambia significativamente los valores.

Veamos los resultados obtenidos en las distintas simulaciones realizadas. En primer lugar, se han fijado los parámetros $a = 10$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$. Para estos, se ha incrementado el valor de la muestra, obteniendo los resultados que se muestran en la Tabla 9.

Tabla 9. Resumen resultados para la ratio y la proporción de cambio.

n	**Media Ratio**	**Desviación típica ratio**	**Media Proporción de cambio**	**Desviación típica Proporción de cambio**
100	1.0004567	0.014226144	0.009904904	0.0048303067
500	0.9998525	0.002478247	0.001938958	0.0008680579
1000	1.0000485	0.001448631	0.001014361	0.0005079548

Fuente: Elaboración propia

Como afirmamos anteriormente, el valor medio de la ratio es muy próxima a 1, y la diferencia disminuye cuando se aumenta el tamaño muestral. Su desviación típica también disminuye al aumentar éste.

De la misma forma, la media de la proporción de cambio tiende a 0 cuando el tamaño aumenta, y su desviación típica también.

Se ha realizado este mismo estudio para tamaños de muestra entre 100 y 5.000, con incrementos de 100. La tabla se ha omitido por su extensión, pero su comportamiento es equivalente, como se muestra en las figuras siguientes.

En la primera de ellas, Figura 10, se muestran los valores medios de las ratios, obtenidas para los distintos valores de n. Como se puede observar, todos se sitúan por debajo de 1.0002 salvo la relativa a $n = 100$, que tiene un valor de 1.00084652. La variación, además se reduce al aumentar la muestra.

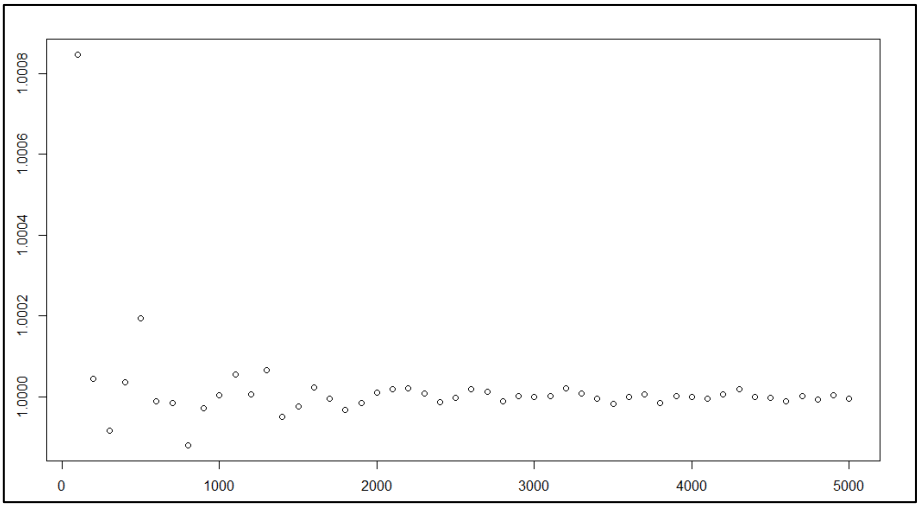


Figura 10. Valor medio de la ratio en función del tamaño muestral.

La disminución de la variabilidad en los valores medios de la ratio puede observarse en la Figura 11, en la que se aprecia un rápido decrecimiento a 0 de la desviación típica al aumentar el tamaño muestral.

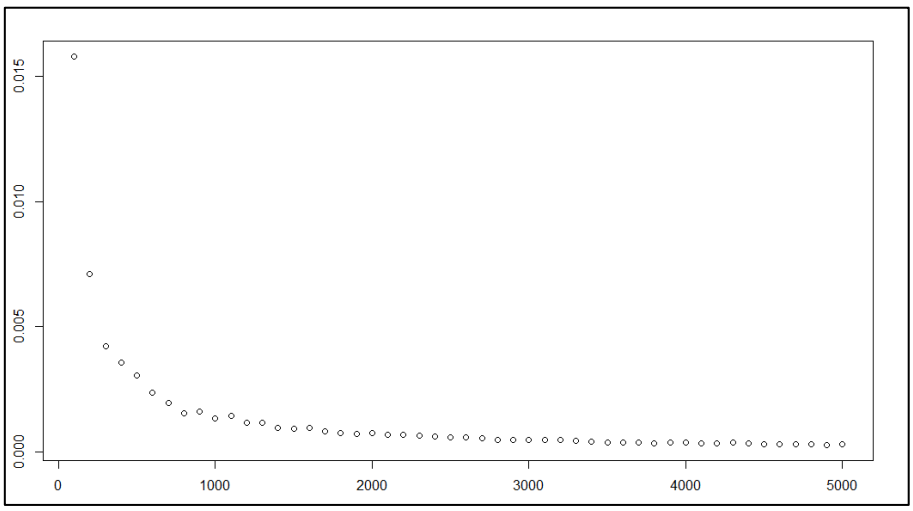


Figura 11. Desviación típica de la ratio en función del tamaño muestral.

En cuanto a la proporción de cambio del valor del criterio de información, su media experimenta un rápido decrecimiento que, al igual que ocurrió en el estudio de la ratio, es superior al resto para $n = 100$, en el que toma el valor 0.01002598. Este valor desciende a menos de la mitad cuando $n = 200$. Este rápido decrecimiento a 0, puede observarse en la Figura 12.

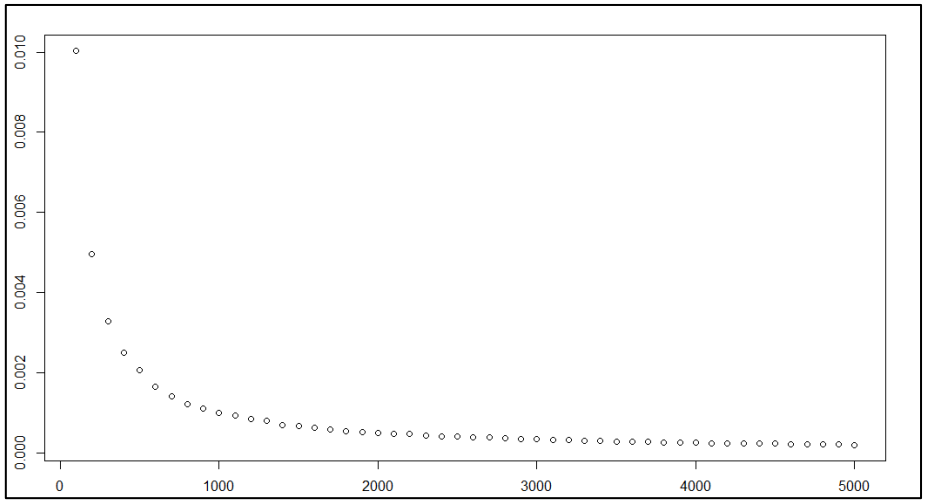


Figura 12. Valor medio de la proporción de cambio en AIC en función del tamaño muestral.

Por último, el comportamiento de su desviación típica es similar al de la media: rápido decrecimiento, con un valor elevado para el valor mínimo del tamaño muestra. Véase la Figura 13.

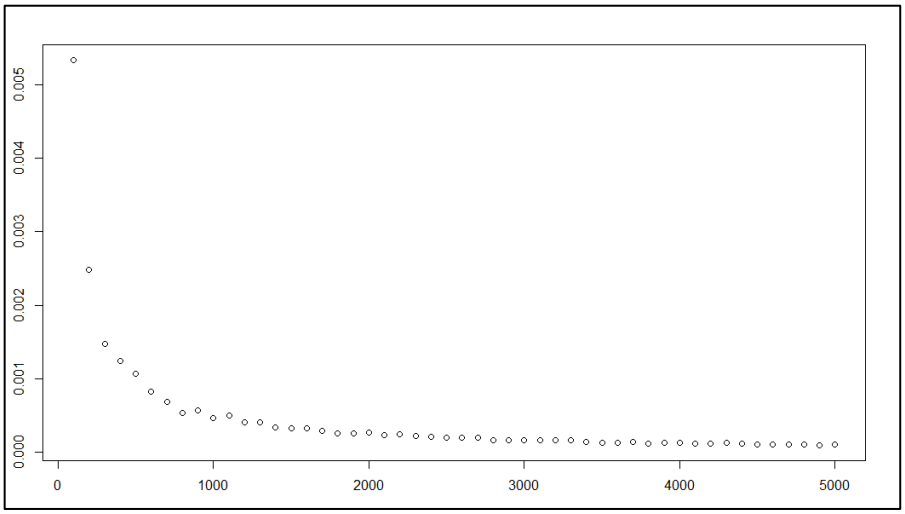


Figura 13. Desv. típica de la proporción de cambio en AIC en función del tamaño muestral.

Veamos, a continuación, el comportamiento de estos parámetros en muestras pequeñas. Para ello, se han realizado nuevas simulaciones para valores de n comprendidos entre 10 y 100. El decrecimiento del valor de la ratio es rápido al aumentar el valor de n como se muestra en la Figura 14. En muestras superiores a 40, el valor se estabiliza en torno a la unidad.

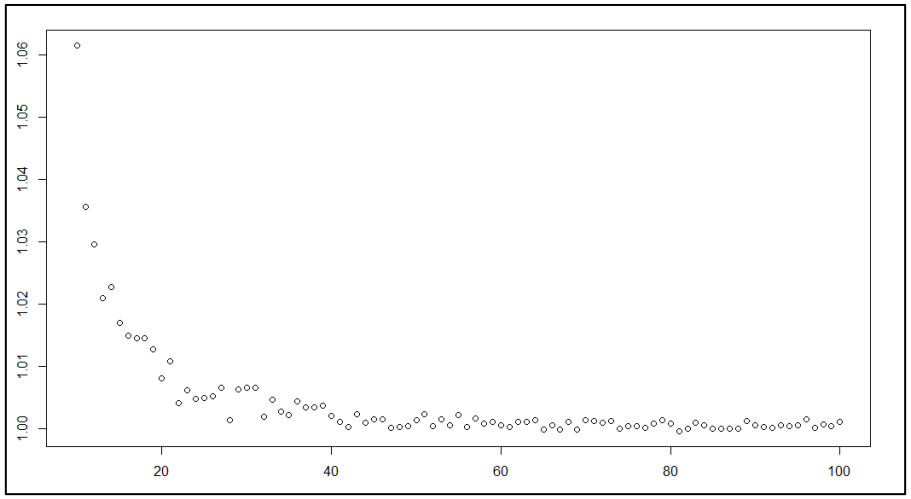


Figura 14. Valor medio de la ratio en función del tamaño muestral. Muestras pequeñas.

No obstante, el decrecimiento de la proporción de cambio es menos acentuado con el tamaño muestral, y comienza a estabilizarse para muestras de mayor tamaño. Este comportamiento se puede ver en la Figura 15.

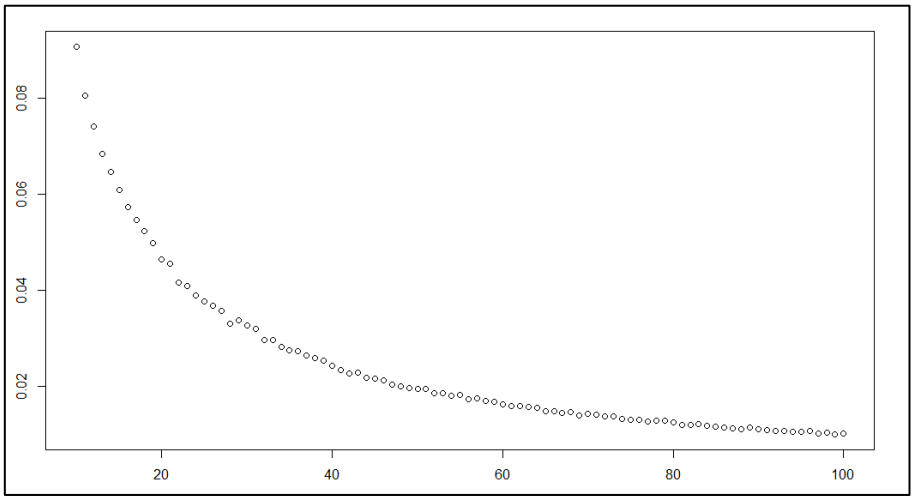
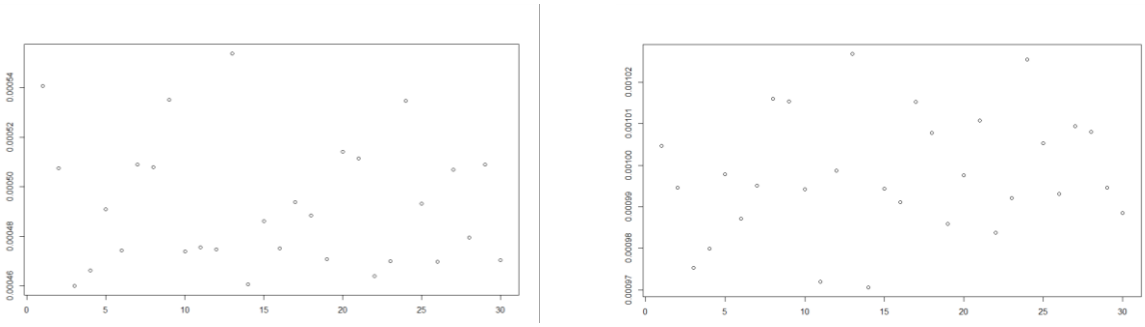


Figura 15. Valor medio de la proporción de cambio en AIC en función del tamaño muestral. Muestras pequeñas.

De todo ello, podemos afirmar que para muestras mayores de 50 la ratio entre las varianzas residuales cuando la observación no es influyente, puede considerarse igual a la unidad. La proporción de cambio se estabiliza a partir de tamaños algo superiores.

En el siguiente grupo de simulaciones, se han fijado los parámetros $n = 1000$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$. El valor de a se ha variado entre 1 y 30. No se ha observado influencia alguna en la ratio ni en la proporción de cambio de AIC que dependa de este valor. La Figura 16, muestra los gráficos de éstos, en función de los valores de a.



Valores de la ratio en función de a | Valores de la prop. de cambio en función de a

Figura 16. Ratio y proporción de cambio en función de a.

Este mismo resultado se ha obtenido en los siguientes casos:

- Se han fijado los parámetros $n = 1000$, $a = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$ y el valor de b se ha variado entre 1 y 30.
- Se han fijado los parámetros $n = 1000$, $a = 10$, $b = 10$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$ y el parámetro que se ha variado es la media de la variable dependiente, también entre 1 y 30, también.
- Se han fijado los parámetros $n = 1000$, $a = 10$, $b = 10$, $\mu_X = 5$ y $s_\varepsilon^2 = 1$ y se ha variado R^2 entre 0.516 y 0.98.
- Se han fijado los parámetros $n = 1000$, $a = 10$, $b = 10$, $\mu_X = 5$ y $R^2 = 0.7$ y el valor de s_ε^2 se ha variado entre 0.5 y 15.

Por lo que podemos afirmar que la proporción de cambio del criterio de información al eliminar una observación no influyente del conjunto de datos es suficientemente pequeña, y el cociente entre sus varianzas es próximo a la unidad para tamaños muestrales no demasiado pequeños.

4.3.3.2. Variación en el caso de observaciones influyentes

El problema que se plantea a continuación es el de determinar el método para la introducción de una observación que pueda considerarse, sin género de dudas, como influyente. Gráficamente, una observación es influyente si la distancia entre la recta de regresión y el punto es significativamente superior a la distancia entre la recta y el resto de puntos.

La distancia vertical a considerar no puede ser considerada en términos absolutos, ya que su magnitud depende, a su vez, de la escala en la que estén expresados los puntos, y, por tanto, de los parámetros escogidos en la simulación.

Por este motivo, para incrementar el nivel de influencia de una observación, se considerará, que para el valor de X situado en punto medio del rango, una componente y, igual a:

$$y_i^* = \hat{y}_i + \Delta\hat{y}_i$$

De forma que, el valor y_i^* varíe entre $\hat{y}_i$ y el máximo valor de la variable Y. No obstante, en rectas con pendiente muy reducida, la desviación del punto respecto a la recta, podría no ser significativa. Para solucionar esto, se considerará $\Delta\hat{y}_i$ con valores comprendidos entre 0 y:

$$max\{y_{max} - \hat{y}_i, x_{max} - x_i\}$$

Y teniendo en cuenta nuevamente la variación de a, b, μ_X, R^2 y s_ε^2, para la creación de las simulaciones se procederá de la siguiente forma:

- Se genera aleatoriamente, con tamaño $n-1$, la variable ε como una distribución Normal con media 0 y varianza residual la seleccionada.
- A partir de la ecuación $s_X^2 = \frac{R^2}{100\cdot(1-R^2)} \cdot \sigma_\varepsilon^2$, con media 10, se genera la variable X como una distribución Normal, nuevamente de tamaño $n-1$.
- Se genera la variable Y a partir del modelo lineal elegido: $Y = a + bX + \varepsilon$. Se estima el modelo de regresión, y se calcula el criterio de información de Akaike, $AIC(-i)$, su varianza residual, $s_\varepsilon^2(-i)$, y la media y la desviación típica de las distancias de los puntos a las rectas.
- A continuación, se introduce la observación n – ésima, que en primer lugar será una observación no influyente, del mismo modo que se realizó en la sección anterior. Esta observación es de la forma $(x, \hat{y})$, con

  $$x = \frac{x_{max} + x_{min}}{2}$$

  e $\hat{y}$ el valor estimado por la recta de regresión creado en el punto 3, más una perturbación aleatoria Normal.
- Con las n observaciones de X e Y, se genera el nuevo modelo para el que se calculará la varianza residual, $s_{\varepsilon N}^2$, el valor del criterio de información, AIC_N, y la

ratio entre $s_{\varepsilon_1}^2$ y $s_{\varepsilon_N}^2$. También, la proporción de cambio entre este modelo y el calculado en el punto 3.

- Por último, para el mismo valor x_i, se considerarán m valores y_i^*, que sustituirán el valor $\hat{y}$ de la observación añadida en el paso 4, por un valor que inicialmente no será influyente, de forma que la distancia vertical entre la observación y la recta sea creciente, por lo que se transformará en una observación que sí será influyente, como se ha comentado anteriormente. Para ello, y_i^* se define de la siguiente forma:
- $y_i^* = \hat{y}_i + \frac{i}{m} max\{y_{max} - \hat{y}_i, x_{max} - x_i\} \quad i = 1, \dots, m$
- Se estima el modelo de regresión del conjunto de datos resultante de añadir esta observación a la muestra de tamaño $n - 1$ creada en los pasos 1 y 2, y se calcula la varianza residual, $s_{\varepsilon I}^2$, el valor del criterio de información, AIC_I, y la ratio entre $s_\varepsilon^2(-i)$ y $s_{\varepsilon I}^2$. Nuevamente, se calcula la proporción de cambio.
- Se considerarán también, valores de y_i^* incluso superiores, de forma que podamos analizar, también, el comportamiento de estos valores cuando la distancia tiende a infinito.

Para llevar a cabo este experimento, en primer lugar, se ha comprobado el aumento experimentado por la ratio de las varianzas residuales y la proporción de cambio de AIC cuando se aumenta la distancia del punto a la recta, es decir, cuando la observación es cada vez más influyente. A continuación, se ha analizado el cambio producido en los resultados al cambiar el tamaño muestral, factor que se ha comprobado antes influyente. Por último se ha analizado la posible variación en los parámetros que determinan el modelo.

Para los valores $n = 1000, a = 10,\ b = 10,\ \mu_X = 5,\ \mathrm{R}^2 = 0.7$ y $s_\varepsilon^2 = 1$, se han considerado 100 observaciones equidistribuidas en el rango descrito. La función que relaciona la distancia entre $\hat{y}_i$ y el valor considerado en cada simulación, con la ratio de varianzas residuales, es función estrictamente creciente, que inicialmente se encuentra ligeramente por debajo de la unidad, 0.9990027, y que toma el valor 1.0258756 en el punto de mayor distancia. Su representación gráfica se puede consultar en la Figura 17.

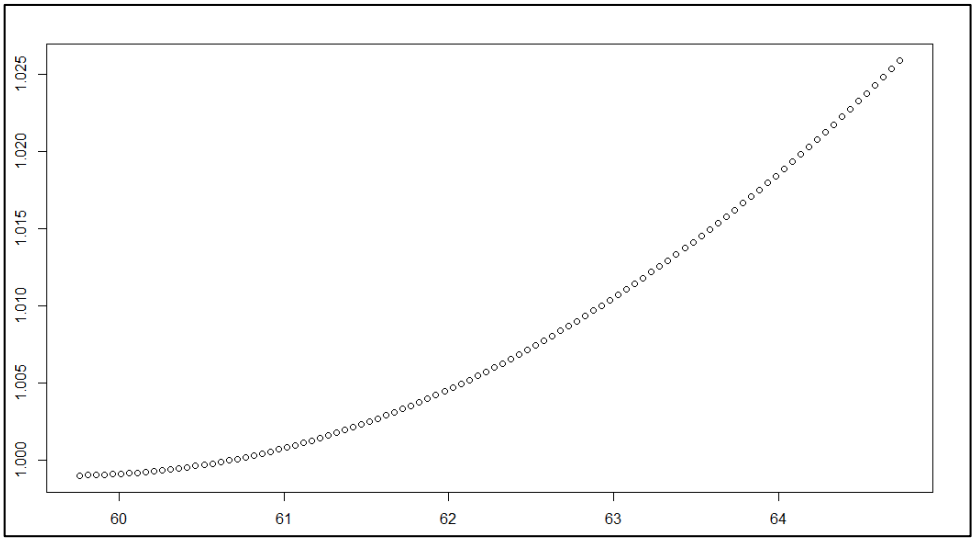


Figura 17. Cambio en la ratio al aumentar la distancia del punto a la recta.

El comportamiento de la proporción de cambio es similar, como puede verse en la Figura 18. De nuevo una función estrictamente con un valor mínimo de ésta de 0.0006391292 y un valor máximo de 0.0100949166, lo que significa un valor casi 16 veces mayor.

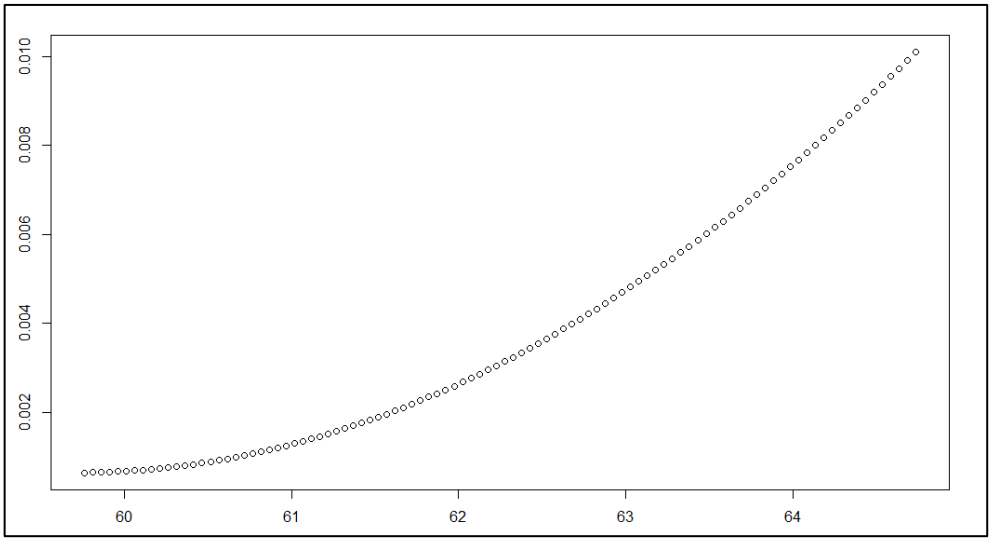


Figura 18. Cambio en la proporción de cambio al aumentar la distancia del punto a la recta.

Es decir, cuando el punto se aleja de la nube de puntos, es decir, se hace influyente, la ratio aumenta alejándose de la unidad, y la proporción de cambio aumenta significativamente, lo que se traduce en un cambio significativo en el valor del criterio de información.

Esto justifica el análisis del cambio de éste para estudiar la existencia de observaciones influyentes. Cabe destacar que, como veremos a continuación, este cambio, aunque en el mismo sentido que el visto, se acelera o se frena en función de algunas características del modelo. Esto invita a identificar las observaciones comparando, entre sí, los valores de AIC de cada uno de los modelos construidos al eliminar una observación del modelo, ya que el cambio producido al eliminar una observación influyente, siempre será significativamente superior que si ésta no lo es.

Veamos a continuación lo que ocurre cuando se incrementa considerablemente la distancia del punto a la recta. En simulaciones realizadas en las que se incrementa el rango de variación de $\Delta\hat{y}_i$, se ha comprobado que el comportamiento de la función que relaciona la distancia con la ratio difiere de la que la relaciona con la proporción de cambio.

La Figura 19 muestra la evolución de la ratio al aumentar la distancia del punto a la recta hasta 50 veces el rango determinado previamente. Este comportamiento no cambia del mostrado en diferentes simulaciones realizadas con hasta 50.000 veces el rango determinado, por lo que podemos afirmar que la ratio tiende a infinito cuando la distancia del punto a la recta también lo hace.

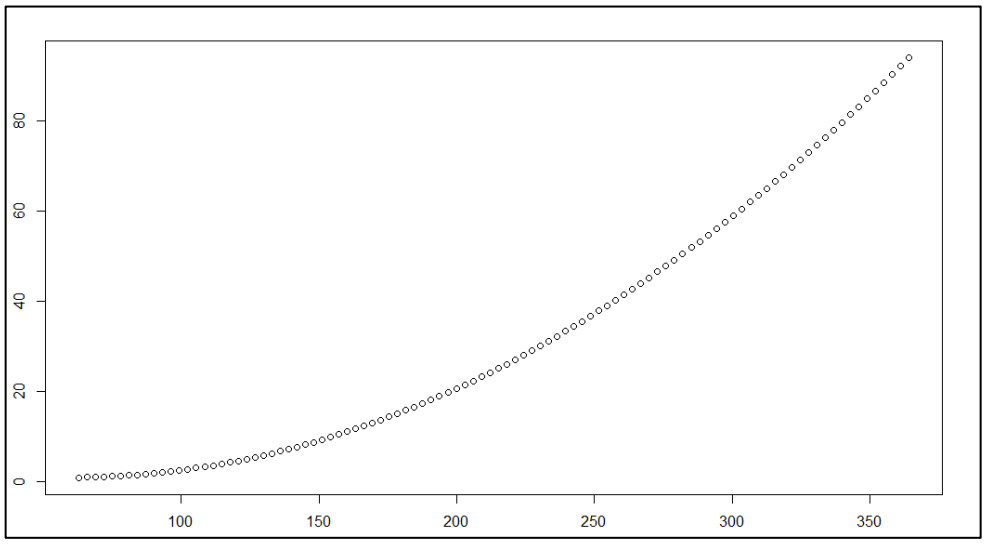


Figura 19. Cambio en la ratio al aumentar significativamente la distancia del punto a la recta.

La proporción de cambio, por el contrario, incrementa su valor con la distancia, al igual que la anterior, pero como cabría esperar por el hecho de ser una proporción, ésta se aproxima al valor 1. Esta convergencia, que puede observarse en la Figura 20 para 50

veces el rango establecido, se ha comprobado realizando simulaciones en las que se han considerado distancias de hasta 10^{150} veces el rango definido.

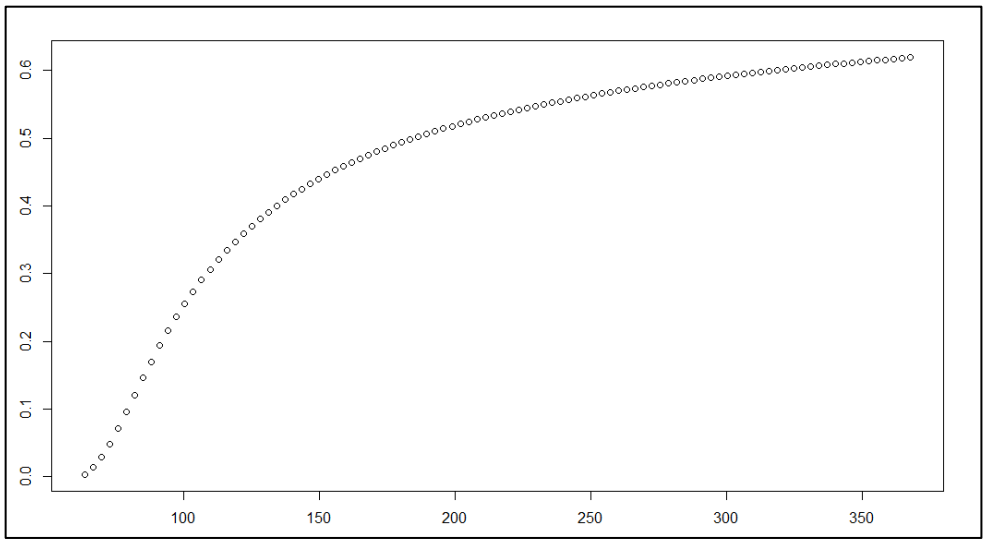


Figura 20. Variación en la proporción de cambio al aumentar significativamente la distancia del punto a la recta.

Mantendremos a partir de ahora como rango de estudio las distancias comprendidas entre 0 y $max\{y_{max} - \hat{y}_i, x_{max} - x_i\}$, como se definió al principio, y analizaremos a continuación como afecta el cambio en el tamaño muestral.

Tabla 10. Valores medios de las ratios y las proporciones de cambio en función de n.

n	**Mínimas alturas medias**	**Máximas alturas medias**	**Mínimas ratios medias**	**Máximas ratios medias**	**Mínimas proporciones de cambio medias**	**Máximas proporciones de cambio medias**
100	60.07276	64.63899	0.9900221	1.210896	0.0063179980	0.07115995
200	60.03600	64.91968	0.9950123	1.118325	0.0032062610	0.04218441
300	60.10896	65.23945	0.9966759	1.089146	0.0021400800	0.03232223
500	60.05560	65.53074	0.9980062	1.060134	0.0012908310	0.02204544
1000	60.01255	65.96450	0.9990037	1.035651	0.0006471602	0.01315847
2000	60.02617	66.21443	0.9995020	1.019248	0.0003240113	0.007166522
5000	60.07807	66.69190	0.9998009	1.008818	0.0001297429	0.003283337

Fuente: Elaboración propia

En la Tabla 10, se muestran los valores medios de 100 simulaciones para cada caso, en las que se han calculado los valores mínimos y máximos de las alturas del punto, y por tanto refleja su distancia, la ratio y la proporción de cambio. Como se ha comentado

anteriormente, al ser, la función que relaciona la distancia con éstos, estrictamente creciente, los valores mínimos y máximos se han alcanzado, en todos los casos, para puntos a la mínima y máxima distancia respectivamente.

Como puede observarse, la ratio media de las observaciones más influyentes disminuye con el tamaño muestral, ya que éste es su comportamiento general, debido a que el aumento en el tamaño muestral hace disminuir el peso que una sola observación tiene sobre el conjunto. Esto mismo ocurre con las proporciones de cambio.

No obstante, sus desviaciones típicas también disminuyen al aumentar n por lo que siempre existen diferencias significativas entre los valores mínimo y máximo. En la se muestra los valores de las desviaciones de las ratios y las proporciones de cambio, así como la probabilidad límite obtenida al realizar contrastes de comparación entre sus valores medios (Tabla 11).

Tabla 11. Desviaciones y contraste de comparación de las ratios y las proporciones de cambio en función de n.

n	**Desviación mínimas ratios**	**Desviación máximas ratios**	**Desviación mínimas proporciones de cambio**	**Desviación máximas proporciones de cambio**	**P – valor comparación de ratios**	**P – valor comparación Proporciones**
100	$8.0243 \cdot 10^{-6}$	0.08024300	0.00018486	0.020577110	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
200	$4.1459 \cdot 10^{-6}$	0.04145902	$6.1516 \cdot 10^{-5}$	0.012067360	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
300	$2.7786 \cdot 10^{-6}$	0.02778581	$3.2027 \cdot 10^{-5}$	0.008396104	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
500	$1.7632 \cdot 10^{-6}$	0.01763177	$1.5250 \cdot 10^{-5}$	0.005555272	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
1000	$7.3424 \cdot 10^{-7}$	0.00734200	$5.7636 \cdot 10^{-6}$	0.002492679	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
2000	$3.5540 \cdot 10^{-7}$	0.00355401	$1.8683 \cdot 10^{-6}$	0.001210974	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
5000	$1.8042 \cdot 10^{-7}$	0.00180418	$5.3523 \cdot 10^{-7}$	0.000628002	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

Fuente: Elaboración propia

De todo ello, y por no apreciar un comportamiento decreciente en el valor del estadístico de contraste, podemos afirmar que, aunque al aumentar el tamaño de la muestra las diferencias se reducen, la variabilidad de sus valores también, que produce, a su vez, un cambio de escala, que mantiene las diferencias significativas.

Veamos a continuación el comportamiento de estos valores al cambiar los parámetros de los que depende el modelo. Comenzaremos por $n = 1000$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$ y a variando entre 1 y 30.

Al variar el término independiente, se produce una traslación vertical de la recta lo que cambia los valores de y_i^*, pero no se aprecia influencia alguna en los valores de la ratio y de la proporción de cambio. Tampoco cuando a toma valores muy elevados.

En la siguiente simulación se tomaron los valores por $n = 1000$, $a = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$ y se varió b entre 1 y 30, lo que cambia la pendiente de la recta. Los resultados fueron similares a los anteriores, con lo que podemos afirmar que tampoco influye el valor del coeficiente del modelo de regresión. Tampoco cuando b toma valores muy elevados.

Una nueva simulación situó como valores constantes $n = 1000$, $a = 10$, $b = 10$, $R^2 = 0.7$ y $s_\varepsilon^2 = 1$ y se consideró valores de μ_X comprendidos entre 1 y 30. Con resultados equivalentes.

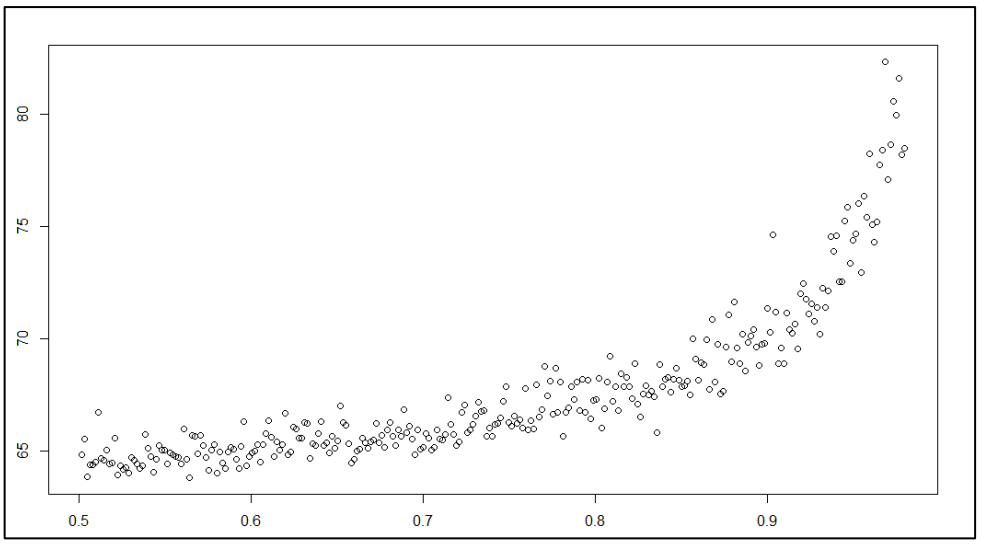


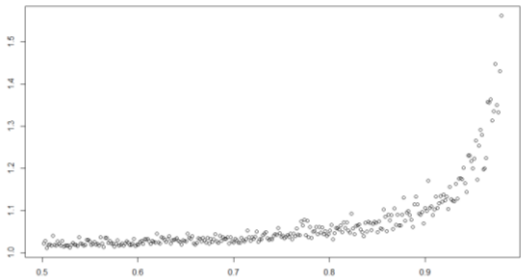
Figura 21. Altura de las observaciones influyentes en función de R^2.

Los resultados cambian cuando varía el valor de R^2, dejando constantes $= 1000$, $a = 10$, $b = 10$, $\mu_X = 5$ y $s_\varepsilon^2 = 1$. Se han realizado 300 simulaciones para valores de R^2 comprendidos entre 0.5016 y 0.9800. La altura de los puntos más alejados aumenta exponencialmente como se observa en la Figura 21.

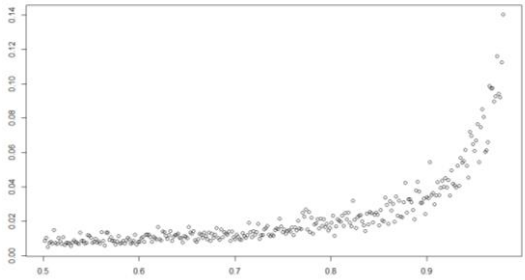
Esto es debido a que, como comentamos anteriormente:

$$s_X^2 = \frac{\mathrm{R}^2}{100(1-\mathrm{R}^2)}\sigma_\varepsilon^2$$

De modo que cuando R^2 tiende a 1, la varianza de X tiende a infinito, y tal y como se ha escogido el rango de variación de las observaciones, su altura, y por tanto la distancia a la recta, aumenta. Este mismo comportamiento también se observa en la ratio y la proporción de cambio de los valores extremos. Véase la Figura 22.



Ratio de los puntos extremos en función de R^2



Proporción de cambio de los puntos extremos en función de R^2

Figura 22. Ratio y proporción de cambio en función de R^2.

Por último, el aumento en la varianza residual genera un aumento, a su vez, en la dispersión de la variable dependiente del modelo debido a su construcción. Esto genera un aumento en la altura de los puntos más alejados de la recta. A su vez, esta mayor dispersión se ve reflejada en la proporción de cambio de los valores más próximos a la recta, que incrementan su valor considerablemente, como puede observarse en la Figura 23, en la que se han analizado los cambios introducidos al variar la varianza residual entre 0.5 y 150. No obstante, este incremento en la proporción de cambio no acerca significativamente los valores de éstos a los más alejados de la recta para valores, que experimentan un leve incremento también. Para comprobar esto último, se han simulado casos con varianzas residuales hasta 10^{10}.

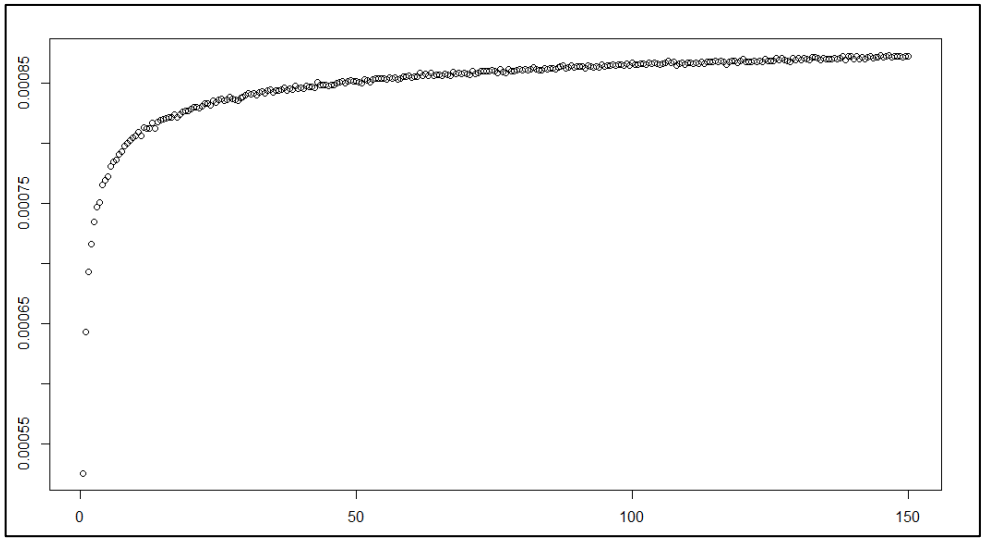


Figura 23. Proporción de cambio de los puntos más próximos a la recta en función de la varianza residual.

4.3.4. Diferencias entre $AIC(-i)$ y $AIC(-j)$

Como se ha comprobado anteriormente, la magnitud de las diferencias obtenidas en el criterio de información al eliminar una observación influyente, puede cambiar cuando se modifica, por ejemplo, el tamaño muestral. Es por ello, que se diseña el método de detección de observaciones influyentes comparando, entre sí, los diferentes valores de AIC resultantes de eliminar una observación de la muestra.

Por tanto, a continuación, se comparará analíticamente los distintos valores del criterio de información obtenidos cuando se elimina una observación i y otra j, de forma que ninguna de ellas sea una observación influyente.

Los valores de los criterios de información de los modelos construidos una vez eliminadas estas observaciones son.

$$AIC(-i) = (n-1)[ln(2\pi)+1] + (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-i)\right) + 2(k+1)$$

$$AIC(-j) = (n-1)[ln(2\pi)+1] + (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-j)\right) + 2(k+1)$$

Por lo que su diferencia es:

$$D_{ij} = AIC(-i) - AIC(-j) = (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-i)\right) - (n-1)ln\left(\hat{\sigma}_\varepsilon^2(-j)\right)$$

Que operando:

$$D_{ij} = AIC(-i) - AIC(-j) = (n-1)ln\left(\frac{\hat{\sigma}_{\varepsilon}^2(-i)}{\hat{\sigma}_{\varepsilon}^2(-j)}\right)$$

Si ninguna de las observaciones es influyente, las varianzas residuales $\hat{\sigma}_{\varepsilon}^2(-i)$ y $\hat{\sigma}_{\varepsilon}^2(-j)$ serán similares, por lo que la diferencia D_{ij} será un valor próximo a 0, que aumentará con el tamaño muestral.

Por todo ello, dado el vector con los criterios de información resultantes de eliminar una observación en cada ocasión, podemos considerar que una observación es influyente si el valor AIC resultante de eliminar dicha observación disminuye significativamente respecto al resto de valores del vector.

Como se analizó en el caso de la diferencia entre el valor del criterio de información de la muestra y el de la muestra a la que se ha eliminado la observación j, veamos la diferencia producida en D_{ij} cuando se modifica el factor de penalización del número de parámetros a estimar.

Para el criterio de información bayesiano, $BIC(-i)$ y $BIC(-j)$ son:

$$BIC(-i) = (n-1)[ln(2\pi) + 1] + (n-1)ln\left(\hat{\sigma}_{\varepsilon}^2(-i)\right) + ln(n-1)(k+1)$$

$$BIC(-j) = (n-1)[ln(2\pi) + 1] + (n-1)ln\left(\hat{\sigma}_{\varepsilon}^2(-j)\right) + ln(n-1)(k+1)$$

Por lo que su diferencia coincide con la calculada para el criterio de información de Akaike.

$$D_{ij} = BIC(-i) - BIC(-j) = (n-1)ln\left(\frac{\hat{\sigma}_{\varepsilon}^2(-i)}{\hat{\sigma}_{\varepsilon}^2(-j)}\right)$$

Al generalizar este factor de penalización al valor P, el resultado es el mismo siempre que éste sea invariante al cambio de muestras con el mismo tamaño, por lo que, aunque éste dependa del tamaño de la muestra, al comparar muestras del mismo tamaño, esta variación no tiene consecuencias sobre el valor de la diferencia.

4.3.5. Desarrollo del método propuesto

Una vez analizados los cambios en el valor del criterio de información de un modelo de regresión, el análisis de estos cambios se realiza de la forma que se muestra a continuación.

Sean $X_1, X_2, \dots, X_k$ un conjunto de k variables explicativas e Y la variable dependiente que se desea estimar a partir del conjunto anterior, para una muestra aleatoria de tamaño n.

El objetivo es detectar la presencia de observaciones influyentes en los datos que pudieran perjudicar la estimación del modelo de regresión.

Para ello, se construyen n muestras de tamaño $n-1$, de forma que la muestra i es la resultante de eliminar la observación i −ésima del conjunto de datos. Para cada una de estas muestras se construye el modelo de regresión para estimar Y en función de $X_1, X_2, \dots, X_k$, y el valor del criterio de información de Akaike de cada uno de ellos.

Una vez hecho esto, obtenemos un vector de valores de AIC que podemos denotar de la siguiente forma:

$$\overrightarrow{\mathbf{AIC}}(-) = \big(AIC(-1), AIC(-2), \dots, AIC(-n)\big)$$

En ausencia de observaciones influyentes, los elementos de este vector serán similares entre sí, ya que todos estos valores se han calculado para muestras que sólo se desvían en una observación que no tiene influencia sobre la estimación del modelo, con el mismo tamaño muestral y con residuos no significativamente diferentes.

Si para una observación i el valor de $AIC(-i)$ del modelo estimado al eliminar dicha observación es considerablemente más pequeño que el del resto de elementos del vector, significará una mejora de su capacidad de ajuste respecto al modelo original, y por tanto la observación i-ésima puede considerarse influyente.

Se propone utilizar dos criterios diferentes para la detección de estas observaciones:

- Considerar una observación atípica cuando su valor $AIC(-i)$ es inferior a la media del vector más de q veces la desviación típica de éste. Es decir:

$$La\ observación\ i\ es\ influyente\ si\ AIC(-i) - E\left(\overrightarrow{\mathbf{AIC}}(-)\right) < q\sigma_{\overrightarrow{\mathbf{AIC}}(-)}$$

- Considerar observaciones influyentes a todas aquellas para las que $AIC(-i)$ esté por debajo de:

$$Q_1 - 3(Q_3 - Q_1)$$

Donde Q_1 y Q_3 son los cuartiles 1 y 3 respectivamente del conjunto de datos. Este método detecta un mayor número de observaciones influyentes debido a que los cuartiles no se ven sesgados como la media y la desviación típica ante la presencia de valores atípicos.

4.3.6. Validez del método. Método de Montecarlo.

Por último, analizaremos, mediante simulación, la efectividad del método, aplicando, a su vez, los dos criterios, y comparando sus resultados con otros métodos propuestos en la literatura y que han sido revisados anteriormente en este mismo capítulo.

Debido a que anteriormente hemos observado que la variación de los coeficientes del modelo, y el aumento en la varianza residual de éste, no resta eficacia a la detección de las observaciones influyentes, se han fijado los parámetros $a = 10$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 5$.

Por lo tanto, para diferentes valores de n, se ha analizado la capacidad de detección de observaciones influyentes del método propuesto, bajo dos criterios, y se ha comparado con los métodos clásicos de la teoría cuando se modifica la distancia del punto a la recta, y que fueron revisados al principio de este capítulo.

Los métodos utilizados, y los criterios para considerar una observación atípica, son los siguientes:

- Método de variación del criterio de información 1: Se identificarán como influyentes aquellos valores inferiores a 3 desviaciones de la media de los $AIC(-i)$.
- Variación del criterio de información 2: Valores por debajo de:

$$Q_1 - 3(Q_3 - Q_1)$$

Valores atípicos

- Valores superiores a 3 desviaciones típicas de los residuos estudentizados del modelo.
- Valores Leverage de las observaciones: Valores por encima de:

$$\frac{2(k+1)}{n}$$

- Método de Bonferroni con probabilidad límite = 0.05.
- DFFITS: Valores mayores que:

$$2\sqrt{\frac{k+1}{n}}$$

- Dfbetas: Valores mayores que:

$$\frac{2}{\sqrt{n}}$$

  en valor absoluto, en cualquiera de sus coeficientes.
- Distancia de Cook: Valores mayores que:

$$\frac{4}{n}$$

- Covratio: Valores por encima de:

$$1+3\frac{k+1}{n}$$

  por debajo de:

$$1-3\frac{k+1}{n}$$

La metodología seguida a continuación es similar a la propuesta anteriormente:

- Se genera aleatoriamente, con tamaño $n-1$, la variable ε como una distribución Normal con media 0 y varianza residual la seleccionada.
- A partir de la ecuación $s_X^2 = \frac{\mathrm{R}^2}{100(1-\mathrm{R}^2)}\sigma_\varepsilon^2$, con media 10, se genera la variable X como una distribución Normal, nuevamente de tamaño $n-1$.
- Se genera la variable Y a partir del modelo lineal elegido: $Y = a + bX + \varepsilon$ y se estima su modelo de regresión.

- A continuación, se introduce la observación n – ésima, en la primera posición en la muestra, para su rápida identificación. En primer lugar, será una observación no influyente, ya que se construye de forma que se sitúe sobre la recta de regresión estimada en el punto 3. Se irá incrementando su altura, de forma que ésta se distancie de la recta, aumentando su capacidad de influencia sobre el modelo. Al igual que se definió anteriormente, esta observación es de la forma $(x, \hat{y})$, con

$$x = \frac{x_{max} + x_{min}}{2}$$

e $\hat{y}$ el valor:

$$y_i^* = \hat{y}_i + \frac{i}{m} max\{y_{max} - \hat{y}_i, x_{max} - x_i\} \quad i = 1, \dots, m$$

- Con las n observaciones de X e Y, se genera el nuevo modelo para el que se analizará, según los nueve métodos propuestos, si la observación es identificada como influyente. Estos resultados serán mostrados en una matriz binaria, en el que 0 indicará que la observación no ha sido considerada influyente por el método, y 1 en caso contrario.

Analizaremos par un valor fijo de n, la sensibilidad de cada uno de los métodos en la detección de la observación influyente, aumentando la distancia del punto a la recta. Como la sensibilidad dependerá de los criterios asignados a cada método, se estudiará con detenimiento la variabilidad en la distancia a la que las observaciones influyentes son detectadas.

Una vez analizado esto, se procederá a estudiar la influencia, que sobre los métodos tiene, el cambio en el tamaño muestral, y a continuación, se analizarán los cambios generados al modificar los parámetros que definen el modelo de regresión.

A su vez, se verá, no sólo si la observación incluida deliberadamente es influyente, también si los métodos detectan otras observaciones que no puedan considerarse de esta forma, y sus resultados cuando se incluye simultáneamente más de una observación de este tipo.

Por último, se analizará la influencia que tiene sobre la capacidad de detección del valor atípico, el que éste se sitúe en lugares alejados de la nube de puntos, pero próximos a la recta de regresión, por lo que no pueden considerarse influyentes. Se aumentará progresivamente la distancia para así aumentar su influencia. Como veremos, esto no afectará significativamente a los métodos basados en la variación de AIC.

Tabla 12. Comparación de métodos de detección de observaciones influyentes. n=1000.

Distanci	AIC	AIC	RESTU	LEVERAG	BONFERRON	DFFI	DFBET	COO	COVRATIO
0,2332	0	0	0	0	0	0	0	0	0
0,4664	0	0	0	0	0	0	0	0	0
0,6997	0	0	0	0	0	0	0	0	0
0,9329	0	0	0	0	0	0	0	0	0
1,1661	0	0	0	0	0	0	0	0	0
1,3993	0	0	0	0	0	0	0	0	0
1,6325	0	0	0	0	0	0	0	0	0
1,8658	0	0	0	0	0	0	0	0	0
2,099	0	0	0	0	0	0	0	0	0
2,3322	0	0	0	0	0	0	0	0	0
2,5654	0	0	0	0	0	0	0	0	0
2,7986	0	0	0	0	0	0	0	0	0
3,0318	0	0	0	0	0	0	0	0	0
3,2651	0	0	0	0	0	0	0	0	0
3,4983	0	0	0	0	0	0	0	0	0
3,7315	0	0	0	0	0	0	0	0	0
3,9647	0	0	0	0	0	0	0	0	0
4,1979	0	0	0	0	0	0	0	0	0
4,4312	0	0	0	0	0	0	0	0	0
4,6644	0	0	0	0	0	0	0	0	1
4,8976	1	0	0	0	0	0	0	0	1
5,1308	1	1	0	0	0	0	0	0	1
5,364	1	1	0	0	0	0	0	0	1
5,5973	1	1	0	0	0	0	0	0	1
5,8305	1	1	0	0	0	0	0	0	1
6,0637	1	1	0	0	0	0	0	0	1
6,2969	1	1	0	0	0	1	0	1	1
6,5301	1	1	0	0	0	1	0	1	1
6,7634	1	1	1	0	0	1	0	1	1
6,9966	1	1	1	0	0	1	0	1	1
7,2298	1	1	1	0	0	1	0	1	1
7,463	1	1	1	0	0	1	0	1	1
7,6962	1	1	1	0	0	1	0	1	1
7,9294	1	1	1	0	0	1	0	1	1
8,1627	1	1	1	0	0	1	0	1	1
8,3959	1	1	1	0	0	1	0	1	1
8,6291	1	1	1	0	0	1	0	1	1
8,8623	1	1	1	0	0	1	0	1	1
9,0955	1	1	1	0	1	1	0	1	1
9,3288	1	1	1	0	1	1	0	1	1
9,562	1	1	1	0	1	1	0	1	1
9,7952	1	1	1	0	1	1	0	1	1
10,0284	1	1	1	0	1	1	0	1	1
10,2616	1	1	1	0	1	1	0	1	1
10,4949	1	1	1	0	1	1	0	1	1
10,7281	1	1	1	0	1	1	0	1	1
10,9613	1	1	1	0	1	1	0	1	1
11,1945	1	1	1	0	1	1	0	1	1
11,4277	1	1	1	0	1	1	0	1	1
11,661	1	1	1	0	1	1	0	1	1

Fuente: Elaboración propia

Iniciamos el conjunto de simulaciones con los parámetros $n = 1000$, $a = 10$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$ y $s_\varepsilon^2 = 5$. Los valores de las distancias verticales o residuos

añadidos a la observación, y las respuestas de los diferentes métodos analizados se muestran en la Tabla 12.

Como puede observarse, el método basado en el cálculo de los COVratios, tiene una sensibilidad para la detección de observaciones influyentes, similar a la obtenida por los métodos propuestos de detección a través de comparación de los AIC. El método más restrictivo en la detección de este tipo de observaciones es el de Bonferroni, y los valores Leverage y los Dfbeta no han logrado identificar la observación influyente. Los métodos basados en el cálculo de Dffits y la distancia de Cook devuelven resultados iguales.

Los valores medios y las desviaciones típicas de las distancias a las que se detecta por primera vez cada uno de los métodos, para el valor definido de los parámetros, en una muestra de 10 simulaciones, en cada una de las cuales se han introducido 100 observaciones individualmente, para muestras de tamaño 1000, se muestra en la Tabla 13.

Tabla 13. *Media y desviación típica de la distancia de la primera observación considerada influyente.*

	Media de distancia	Desviación típica de distancia
AIC1	5,13040	0,15316885
AIC2	5,10543	0,15249881
Residuos estudentizados	6,77637	0,16040036
Leverage	-	-
Bonferroni	9,18922	0,18104320
DFFITS	6,29967	0,25314897
DFBETAS	-	-
Distancia de Cook	6,29967	0,25314897
Covratio	4,84278	0,12672752

Fuente: Elaboración propia

La distancia media menor, y por tanto el método más sensible es el del cálculo de los COVratios, seguido de los métodos propuestos. Este es también el comportamiento de las desviaciones típicas de las distancias.

Los métodos Leverage y Dfbetas no han detectado, en ningún caso, la observación como influyente. En una de las simulaciones realizadas, la probabilidad límite de

Bonferroni ha identificado como influyentes a todas las observaciones propuestas. Lo que ha reducido la distancia a la que se detecta ésta a 0.2414, valor que se ha eliminado del cálculo de los estadísticos.

En el siguiente grupo de simulaciones, se ha ampliado el rango de variación de y_i^* para determinar el grado de sensibilidad de los métodos que no han logrado la detección de observaciones influyentes.

Tras esto, Dfbeta ha detectado observaciones influyentes a partir de distancias en el entorno de los 20 puntos. No se ha logrado la detección, por parte de Leverage, en las distancias comprobadas, menores o iguales a 1000.

A continuación, se han realizado tres grupos 20 de simulaciones, para $n = 100$, $n = 200$ y $n = 500$. En cada una de ellas se han introducido, individualmente, 50 observaciones en el rango propuesto y ordenadas en distancia creciente. Las distancias mínimas a las que se han detectado con cada método y con cada tamaño muestral, se puede ver en la Tabla 14.

Tabla 14. Media y desviación típica de la distancia de la primera observación considerada influyente. Diferentes tamaños muestrales.

	n=100		n=200		n=500	
	Media de distancia	**Desv típica de distancia**	**Media de distancia**	**Desv típica de distancia**	**Media de distancia**	**Desv típica de distancia**
AIC1	5,0895	0,4914	5,2130	0,2282	5,2805	0,2401
AIC2	5,0709	0,5698	5,0988	0,2688	5,0966	0,1941
Restud	6,7322	0,5352	6,8610	0,2604	6,8359	0,2506
Leverage	-	-	-	-	-	-
Bonfer	7,9074 (17/20)	3,0679	8,3785 (18/20)	0,2548	8,8702 (19/20)	0,3277
DFFITS	6,0932	0,5182	6,1910	0,3309	6,14656	0,4558
DFBETAS	12,2367 (3/20)	4,7282	8,4855 (3/20)	1,2993	8,05596 (5/20)	1,8735
Cook	6,3083	0,5657	6,3271	0,3549	6,169215	0,4514
COVratio	4,7739	0,3549	4,8862	0,1930	4,869635	0,1729

Fuente: Elaboración propia

No existen diferencias significativas en las distancias a las que los métodos detectan las observaciones influyentes, salvo en el método de Bonferroni, en el que ésta aumenta y Dfbeta en la que parece disminuir.

Se aprecia una disminución en la desviación típica en los métodos AIC1, AIC2, residuos Estudentizados y Covratio. El método Dfbeta sólo detecta observaciones influyentes en 3, 3 y 5 simulaciones de cada 20 realizadas y el método de Bonferroni en 17, 18 y 19, respectivamente.

De todas las simulaciones realizadas, pueden descartarse los métodos del cálculo de valores Leverage y Dfbeta, el primero por la imposibilidad de detectar observaciones influyentes, y el segundo por su reducida sensibilidad, que podría solucionarse reduciendo el valor límite impuesto, y por su elevada variabilidad en la detección.

El método de Bonferroni ha identificado, en algunas simulaciones, todas las observaciones como influyentes, lo que es, evidentemente, erróneo; y no ha identificado ninguna en otras, en las mismas condiciones en las que sí lo había hecho previamente. Por tanto, tiene un buen comportamiento en general, pero en ocasiones no funciona correctamente.

Los métodos basados en el cálculo de Dffits y de la distancia de Cook, con resultados muy similares, han arrojado buenos resultados en cuanto a homogeneidad en la detección de observaciones influyentes, pero en menor medida que el método del cálculo de los COVratios, los basados en AIC y en el cálculo de valores atípicos de los residuos Estudentizados. Los tres primeros, con resultados similares, y el cuarto con menor sensibilidad.

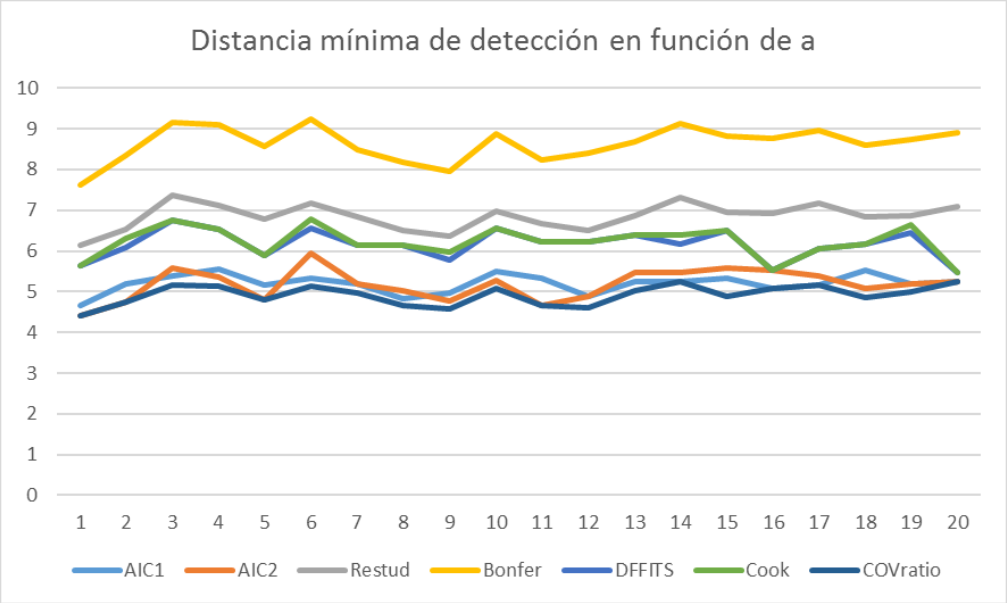


Figura 24. Distancias mínimas de detección de observaciones influyentes en función de a.

Analicemos ahora, para $n = 300$, $b = 10$, $\mu_X = 5$, $R^2 = 0.7$, $s_\varepsilon^2 = 5$ y a tomando valores enteros entre 1 y 20. La Figura 24, resume la información obtenida relativa a la distancia a la que cada método ha detectado la primera observación influyente. Se han eliminado del estudio los métodos Leverage y Dfbeta.

Como se puede observar, el incremento del valor de a no cambia la distancia mínima a la que la observación es detectada como influyente por cada método.

Para $n = 300$, $a = 10$, $\mu_X = 5$, $R^2 = 0.7$, $s_\varepsilon^2 = 5$ y b entre 1 y 20 se observa el mismo comportamiento independiente del valor de b. De nuevo, se resume la información gráficamente, en la Figura 25. De ambos resultados, podemos afirmar que los métodos no son sensibles al cambio en los coeficientes del modelo.

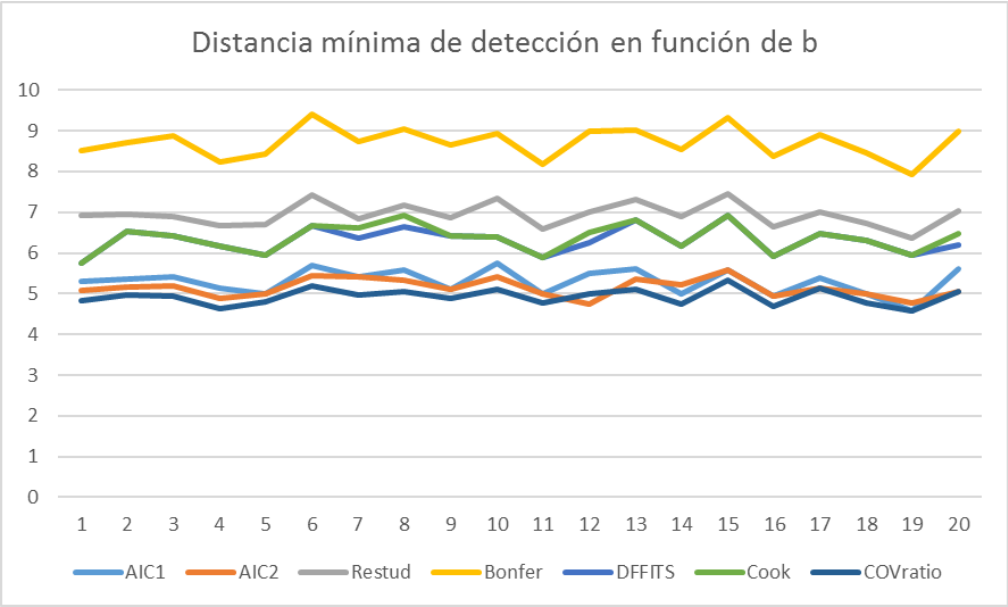


Figura 25. Distancias mínimas de detección de observaciones influyentes en función de b.

El cambio en el valor de la media de la variable independiente tampoco produce cambios significativos en la sensibilidad de la detección de las observaciones en función de su distancia. Puede verse en la Figura 26 el resultado de las 20 simulaciones realizadas en las que se ha variado el valor de μ_X desde 1 hasta 20.

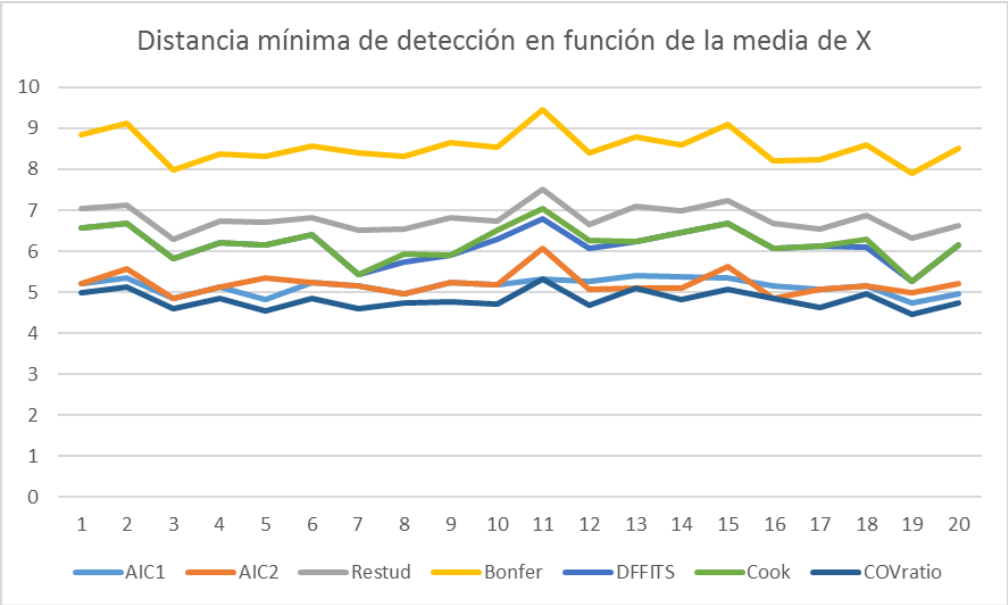


Figura 26. Distancias mínimas de detección de observaciones influyentes en función de la media de X.

El valor de R^2 tampoco influye en la detección de las observaciones como puede apreciarse en la Figura 27. Hemos comentado previamente de la falta de detección, en

ocasiones, de observaciones influyentes por parte del método de Bonferroni. En este caso, se han dado dos casos, en los que esto ha ocurrido.

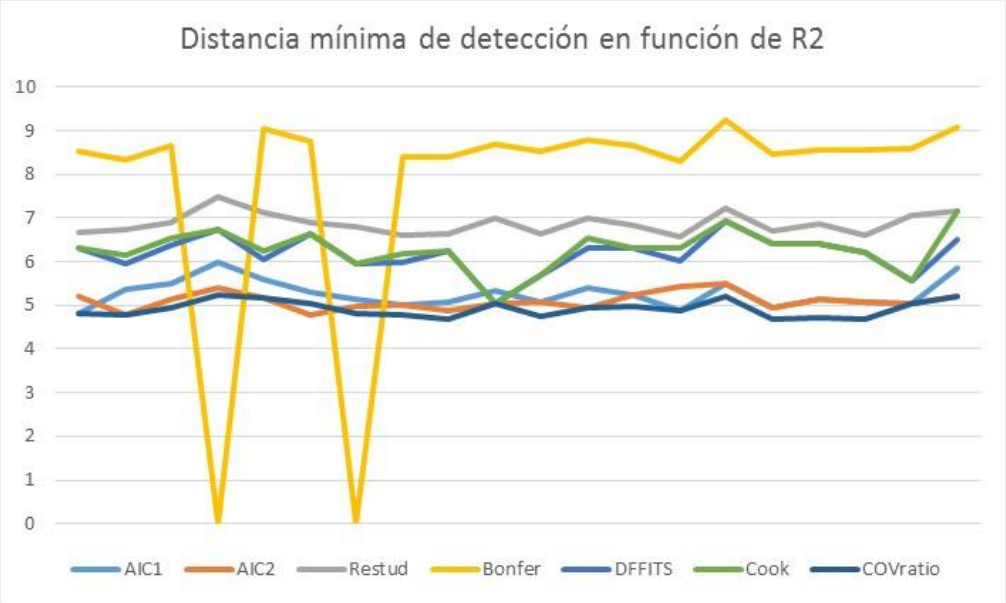


Figura 27. Distancias mínimas de detección de observaciones influyentes en función R^2.

No obstante, el último parámetro analizado, la varianza residual, sí que ha revelado influir significativamente sobre la efectividad de los métodos analizados. Esto no debe sorprender debido a que, al incrementar la varianza residual, aumenta la dispersión de los puntos y por tanto se incrementa la distancia necesaria para considerar una observación influyente. La muestra el crecimiento de la distancia mínima de detección en función del valor de la varianza residual, entre 0.5 y 10. El incremento experimentado por el método de Bonferroni es superior al resto. (Figura 28)

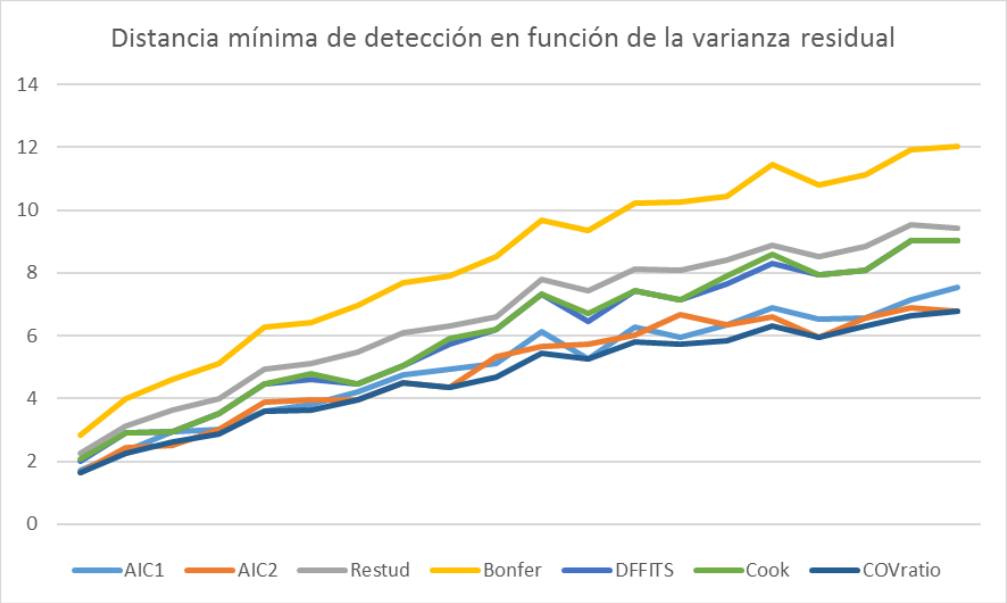


Figura 28. Distancias mínimas de detección de observaciones influyentes en función de la varianza residual.

Para terminar, veamos la respuesta de los métodos a la introducción de diferentes observaciones influyentes de forma simultánea, observando también la detección de otras observaciones, además de las incluidas.

4.3.6.1. Otras simulaciones. Observación influyente en la cola inferior de la distribución

En la primera simulación realizada se ha introducido, para el percentil 5 de la variable X, un valor de Y igual a su valor medio. El resumen de resultados se muestra en la Tabla 15.

Todos los métodos propuestos identifican correctamente la observación influyente salvo los valores Leverage que se centran exclusivamente en las observaciones incluidas en los extremos de la nube de puntos.

Tabla 15.Resultado de la detección de la observación influyente en P_5.

Tamaño	**1000**		
Variable dependiente	$X \sim N(10, 5)$		
Modelo de regresión	$Y = 3 + 2 \cdot X + N(0, 1)$		
Observación influyente: Percentil de *X*	5		
Observación influyente: Valor *Y*	Media		
Observaciones detectadas	Propuesta	Otras	Con el valor extremo
AIC: Método de tipificación	☒	☐	☒
AIC: Método de los cuartiles	☒	☒	☒
Residuos estudentizados: tipificación	☒	☐	☒
Valores leverage de las observaciones	☐	☒	☐
Método de Bonferroni	☒	☐	☒
Valores DFFITS	☒	☒	☒
Coeficientes DFBETAS	☒	☒	☒
Valores de distancia de Cook	☒	☒	☒
COVratio	☒	☒	☒

Fuente: Elaboración propia

Cabe destacar, que los únicos métodos que han identificado, en todos los casos, únicamente la observación propuesta como influyente han sido los métodos de tipificación de los criterios de información y de los residuos estudentizados; y el método de Bonferroni. El resto, han identificado, además, otras observaciones como atípicas, en algunos casos, un número elevado de ellas. Estas otras observaciones detectadas se concentran, para los valores Leverage y los COVratios, en la mayor parte de los casos, en los puntos extremos de la nube, que aun pudiéndose considerar atípicos, no pueden ser vistos como influyentes. No obstante, las observaciones, distintas a la propuesta, detectadas por el método de los cuartiles para los valores AIC, no se sitúan en los extremos de la nube. Es más, no detecta observaciones influyentes en los extremos de la nube, salvo en los casos en los que realmente son de tal tipo.

Las observaciones detectadas por el método DFFITS y el del cálculo de la distancia de Cook son coincidentes en casi todas las simulaciones. Además, una gran mayoría de

las observaciones cercanas a los extremos, y detectadas por el método de los cuartiles de los AIC, son también detectadas por estos métodos.

La detección de la observación influyente en el percentil 5, se ha realizado, además, con el valor extremo de cada método, en todos los casos. También en el caso de los métodos en los que se calcula el criterio de información, en los que el valor de $AIC(-50)$ es el mínimo de todos los incluidos en el vector $\overrightarrow{\mathbf{AIC}}(-)$.

En la Figura 29 se muestra el gráfico de dispersión y la observación influyente propuesta rodeada en un círculo, que puede identificarse claramente como influyente. Además, se ha recuadrado un ejemplo de observación detectada por el método de los valores Leverage y COVRATIO.

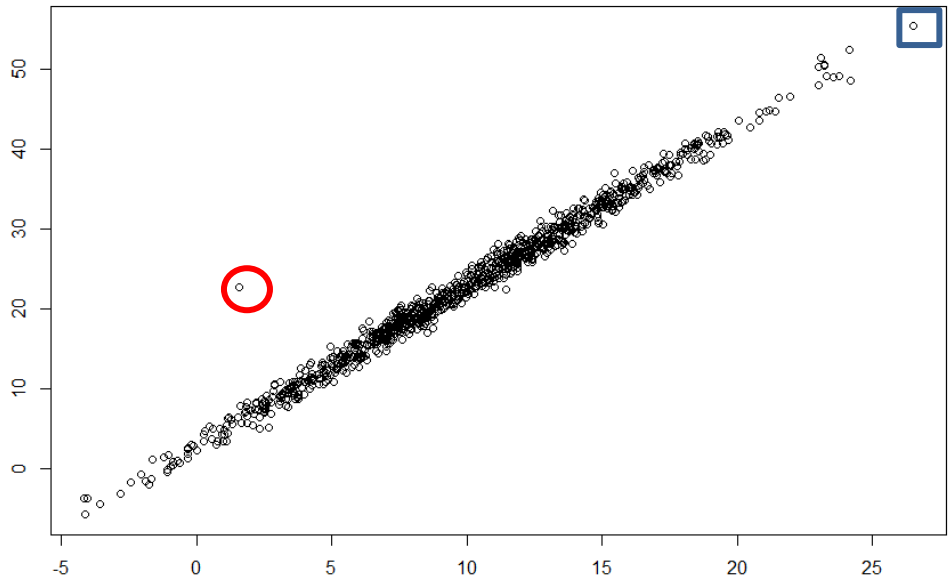


Figura 29. Gráfico de dispersión de la observación influyente en P_5.

En todas las simulaciones realizadas, salvo aquellas en las que la observación se introdujo en el extremo exacto de la nube, el método del cálculo de valores Leverage, no fue capaz de identificar dicha observación, por lo que, aunque en la simulación se siguió considerando, dejarán de significarse sus resultados a partir de este momento.

4.3.6.2. Otras simulaciones. Dos observaciones influyentes en el centro de la distribución

A los dos valores centrales de la variable X, es decir, $X_{(500)}$ y $X_{(501)}$, se les han asignado, en el siguiente conjunto de simulaciones, el valor mínimo y máximo de Y respectivamente, de forma que se generan dos observaciones influyentes, como se observa en la Figura 30.

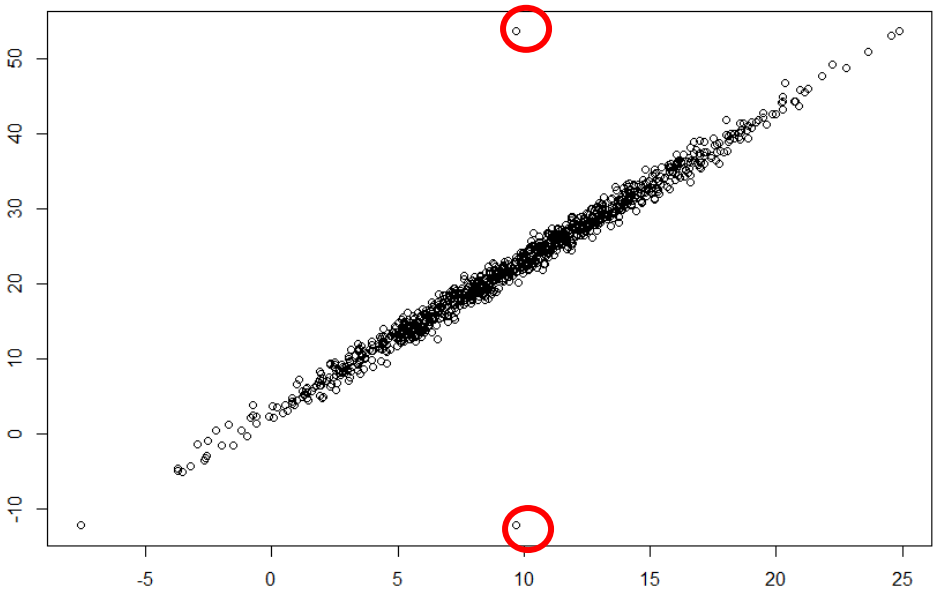


Figura 30. Gráfico de dispersión con 2 observaciones influyentes en los valores centrales. (Fuente: Elaboración propia.)

De nuevo, todos los métodos los detectan con sus valores extremos, unas veces una y otras veces la otra observación, y los métodos de tipificación de AIC y de los residuos estudentizados, así como el método de Bonferroni, los detectan en exclusiva.

Los resultados obtenidos con el cálculo de los DFFITS y las distancias de Cook coinciden en todos los casos, y los COVRATIOS detectan, en todos los casos, un mayor número de observaciones, fundamentalmente, de nuevo, en los extremos.

4.3.6.3. Otras simulaciones. Diversas observaciones influyentes en el centro de la distribución

A continuación, introducimos un mayor número de observaciones influyentes, para detectar la sensibilidad de los distintos métodos. En primer lugar, introducimos para los percentiles 5, 10, 15, 85, 90, y 95 de la variable X, un valor de Y igual a su media. De esta forma, analizaremos, también, la introducción de observaciones influyentes desplazadas en el eje X.

También, en este caso, todos los métodos detectaron las seis observaciones influyentes. Los valores extremos de los métodos se alcanzaron en unas ocasiones en el percentil 5, y en otras en el 95. Los resultados son, de nuevo, similares a los obtenidos anteriormente: Tres métodos: tipificación de AIC, tipificación de residuos Estudentizados y Bonferroni, detectan únicamente las seis observaciones impuestas, el método de los cuartiles de AIC que detecta, además, otras observaciones a lo largo de todo el rango de la variable, y DFFITS, DFBETA, distancia de Cook y COVRATIO, en mayor medida, que además de las seis observaciones, incluye otras que están situadas, en casi todos los casos, en los extremos de la nube de puntos.

Los resultados de la distancia de Cook y de COVRATIO coinciden en el 95 % de las simulaciones realizadas.

4.3.6.4. Otras simulaciones. Observaciones influyentes y no influyentes en el centro de la distribución

A la vista de que todas las observaciones introducidas han sido reconocidas por todos los métodos, la introducción de un conjunto de observaciones con distinto grado de influencia, junto con observaciones no influyentes es el siguiente paso medir la sensibilidad. Para ello, se ha asignado en los percentiles 5, 10, 15, 20, …, 95 de la variable X, el valor medio de Y.

En la Figura 31 se muestran las 19 observaciones introducidas en color rojo para distinguirlas del resto de la nube de puntos. Es evidente, que las que están situadas en los

extremos son observaciones influyentes, mientras que las centrales no lo son. Veamos, para los métodos estudiados, qué observaciones son identificadas como tales.

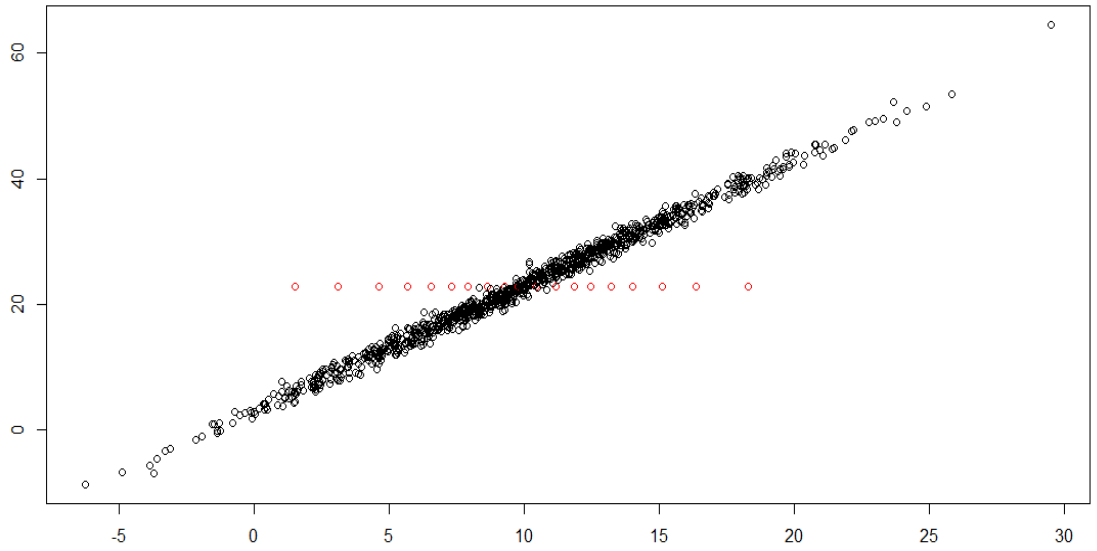


Figura 31. Gráfico de dispersión con observaciones en P_{5k}.

La Tabla 16 resume los resultados obtenidos en las 100 simulaciones realizadas, indicando en cuantas ocasiones, cada uno de los métodos ha detectado las observaciones como influyentes.

Las tres observaciones centrales, correspondientes a los percentiles 45, 50 y 55 y que no se distinguen en el gráfico de dispersión de la Figura 31 por estar integrados en la nube de puntos, son identificados como valores no influyentes por todos los métodos estudiados.

Las observaciones situadas en los percentiles 40 y 60 sólo son identificadas, en un 87 % y un 85 % de las ocasiones, por el método de los cuartiles de AIC, mientras que el resto de métodos nunca las considera de esta forma.

El método de los cuartiles de AIC considera, en el 100 % de los casos, que el resto de observaciones son influyentes.

Además, todos los métodos coinciden, en el 100 % de las ocasiones, que los percentiles 5, 10, 85, 90 y 95 se corresponden con observaciones influyentes. Los percentiles 15 y 80 también son considerados de esta forma en todos los casos, salvo en

el 4 % y el 3 % de las veces en las que se ha aplicado el método de tipificación de los residuos Estudentizados.

El método de tipificación de los AIC considera que las observaciones comprendidas entre el percentil 25 y el 75 no son influyentes salvo en el 11 % de las ejecuciones, para la primera, y el 4 % para la segunda.

Tabla 16. Resultado de la detección de las observaciones en P_{5k}.

Percentil	**Método**							
	AIC 1	**AIC 2**	**Resid. est.**	**Bonf.**	**DFFITS**	**DFBETAS**	**COOK**	**COVRATIO**
P_5	100	100	100	100	100	100	100	100
P_{10}	100	100	100	100	100	100	100	100
P_{15}	100	100	96	100	100	100	100	100
P_{20}	100	100	0	100	100	100	100	100
P_{25}	11	100	0	96	100	100	100	100
P_{30}	0	100	0	0	100	100	100	100
P_{35}	0	100	0	0	18	46	17	99
P_{40}	0	87	0	0	0	0	0	0
P_{45}	0	0	0	0	0	0	0	0
P_{50}	0	0	0	0	0	0	0	0
P_{55}	0	0	0	0	0	0	0	0
P_{60}	0	85	0	0	0	0	0	0
P_{65}	0	100	0	0	19	0	16	98
P_{70}	0	100	0	0	99	9	99	100
P_{75}	4	100	0	95	100	100	100	100
P_{80}	100	100	97	100	100	100	100	100
P_{85}	100	100	100	100	100	100	100	100
P_{90}	100	100	100	100	100	100	100	100
P_{95}	100	100	100	100	100	100	100	100

Fuente: Elaboración propia

Por tanto, los tres métodos que menos observaciones identifican son los de tipificación de los residuos, y de los valores AIC; y el de Bonferroni. En el lado contrario, los que más observaciones detectan son el del cálculo de los cuartiles de AIC y el de COVRATIO, aunque éste último es el que más observaciones adicionales incluye, la mayor parte de ellas, en los extremos.

Los resultados de DFFITS y las distancias de Cook, respecto a la detección de las observaciones propuestas, coinciden en el 97 % de las simulaciones.

4.3.6.5. Otras simulaciones

Cuando no se incluyen observaciones influyentes de forma deliberada, los dos métodos en los que participan los valores AIC, dan resultados prácticamente iguales, detectando un número de observaciones comprendido entre 8 y 17 observaciones influyentes. Los métodos DFFITS, DFBETA, distancia de Cook y COVRATIO detectan un gran número de observaciones influyentes (entre 40 y 60). El método de tipificación de los residuos estudentizados no detecta, en ningún caso, ninguna observación influyente, y el método de Bonferroni detecta una única observación de este tipo, en todos los casos.

Cuando se aumenta la varianza residual, y, por tanto, disminuye la capacidad de ajuste del modelo, se añaden al método de tipificación de AIC, otras observaciones adicionales a las propuestas, pero no en el método de Bonferroni ni en el de tipificación de los residuos estudentizados. Este último método, es, además, en el que más rápidamente disminuye el número de observaciones detectadas. Es más, en algunas simulaciones no se detectaron observaciones influyentes con una desviación típica residual de 3.

Como cabía esperar, el número de percentiles identificados como influyentes entre los propuestos, disminuye en todos los métodos al aumentar la varianza.

Variaciones en el signo de la pendiente y en la ordenada en el origen no generan cambios significativos en los resultados.

4.3.7. Conclusiones obtenidas a partir de las simulaciones realizadas

El criterio de información es una medida relativa de la bondad de ajuste del modelo. Al disminuir este valor, el ajuste de la función calculada para un determinado conjunto de datos, mejora. Dada una observación influyente sobre un conjunto de datos determinado, si eliminamos de éste, dicha observación, el modelo estimado sobre el

conjunto de datos resultante tendrá un mejor ajuste que el primero, por lo que el valor de su criterio de información será menor al del inicialmente calculado.

El método analizado evalúa la presencia de observaciones influyentes midiendo la variación que cada observación, al ser eliminada del conjunto, produce en el valor de AIC. En todas las simulaciones realizadas se ha obtenido que la variación de este valor, y de los elementos que lo contienen, la variación es significativa cuando la observación es influyente, mientras que no lo es cuando no puede considerarse la observación de este tipo.

Por todo ello, podemos afirmar que la diferencia entre el valor del criterio de información de un modelo es significativamente mayor que el del modelo estimado a partir del conjunto de datos al que se ha extraído la observación influyente. También, si la observación no es influyente, la proporción de diferencia producida será despreciable.

Por tanto, y para comparar valores del criterio de información para conjuntos de datos prácticamente iguales, con el mismo número de observaciones, se comparan los valores de los criterios de información de todos los conjuntos de datos resultantes de eliminar una observación.

Se han propuesto dos variantes del método, la primera usando la desviación respecto al valor mediano de la variable, y tomando como medida de la desviación, el rango intercuartílico, y la segunda, respecto al valor medio, en el que se utilizó la desviación típica como medida de dispersión.

La primera de las variantes es menos sensible a la introducción de observaciones influyentes, y por tanto detecta un mayor número de ellas; y la segunda tiende a desplazar su "centro de masas" hacia el extremo inferior del conjunto de valores AIC, por lo que es más exhaustivo en la detección de observaciones influyentes.

En todas las simulaciones (miles de ellas), sin excepción, se han detectado las distintas observaciones influyentes, en las diferentes situaciones propuestas, con las dos variantes del método propuesto. Además, se han comparado dichos métodos con otros propuestos anteriormente y que son clásicos en la bibliografía relativa a los valores atípicos, con grandes resultados, ya que, con éste, se obtienen los mismos resultados para las

observaciones influyentes, y, al contrario que los métodos tradicionales, éste no identifica como tales, observaciones que, aun siendo atípicas, no ejercen influencia alguna en la estimación del modelo.

Por tanto, podemos afirmar que los métodos basados en la variación de AIC, aunque costosos computacionalmente, son eficaces para detectar valores atípicos que son, además, influyentes para el modelo de regresión, diferenciándolas de las que no lo son.

4.3.8. Método de selección para la construcción del modelo

La construcción de un modelo de regresión, tiene como objetivo explicar el comportamiento de una variable, la dependiente, en función de un conjunto de variables, capaces de explicarlo. En la metodología de precios hedónicos, estas variables explicativas son las características que definen un bien heterogéneo, y la variable explicada es el precio de bien.

Para un conjunto de datos dado, puede existir un subconjunto de observaciones para las que el precio del bien no se explica de la misma forma, a partir del valor de sus características, que el resto de observaciones.

Si el objetivo es explicar, de forma general, el comportamiento del precio de un bien, en función del valor de las características que lo definen, y estas observaciones desvirtúan esta explicación por su distinto comportamiento, el investigador puede decidir eliminarlas del conjunto de datos en pro de construir un buen modelo que explique el resto de observaciones.

La eliminación de observaciones de la muestra puede generar un sesgo en la construcción del modelo que debe valorar el investigador. A continuación, se propone una metodología iterativa, a partir del método ya propuesto y basado en el cálculo del valor del criterio de información, para la selección de las observaciones que deben ser eliminadas del conjunto de datos porque no pueden ser explicadas a través del mismo proceso que el resto.

Es necesario realizar tres apreciaciones importantes antes de la explicación del método.

- El método devuelve el conjunto de observaciones que deben ser eliminadas por su alta influencia sobre el modelo, pero el investigador, es, en última instancia, el que debe decidir, una a una, la pertinencia de su eliminación, debido al sesgo que se introduce en el modelo.
- El modelo construido con el conjunto final de datos tiene un ajuste mejor, como cabe esperar, que el modelo inicial y, por tanto, su coeficiente de determinación será significativamente mejor.
- El método es computacionalmente muy costoso, sobre todo en muestras de gran tamaño, y su convergencia está pendiente de ser demostrada, aunque todas las simulaciones realizadas han resultado positivamente convergentes.

Los pasos para la aplicación del método de selección de observaciones son los siguientes:

- En primer lugar, se construye, a partir de un conjunto de datos dado, el modelo de regresión deseado y se evalúa la capacidad de ajuste del modelo.
- A continuación, se elimina la primera observación del conjunto de datos, se estima el nuevo modelo a partir del conjunto de $n-1$ observaciones resultantes, y se calcula el valor del criterio de información del nuevo modelo $AIC(-1)$. Este proceso se repite para cada una de las observaciones que conforman la muestra, obteniendo así, un vector de valores de criterios de información: $\boldsymbol{AIC}$.
- Se evalúa la existencia de valores atípicos en la cola inferior del vector, que puedan identificarse como observaciones influyentes. Esto puede realizarse con los criterios propuestos, esto es, se consideran atípicos del vector, a aquellos valores $AIC(-i)$ que verifican:

$$AIC(-i) - E\left(\overrightarrow{\mathbf{AIC}}(-)\right) < k\sigma_{\overrightarrow{\mathbf{AIC}}(-)}$$

o

$$AIC(-i) < Q_1 - k(Q_3 - Q_1)$$

- En caso de que exista elementos del vector considerados atípicos, se selecciona el valor menor de ellos, para eliminarlo de la muestra. Una vez eliminado, se inicia el procedimiento desde el paso 1, con la muestra, de $n-1$ elementos, obtenida.

El proceso se repite hasta que ningún elemento del vector sea considerado atípico.

El número de observaciones eliminadas de la muestra por ser influyentes, depende del valor de k impuesto en la identificación de valores atípicos.

5.Análisis de los principales portales inmobiliarios de España

En el presente capítulo se analizan los portales inmobiliarios más importantes en España, en la actualidad, teniendo en cuenta criterios como el número de inmuebles ofertados a la venta, la estructura y la usabilidad de la página, el número de visitantes únicos al día del portal, el ranking Alexa, la existencia de aplicación para dispositivos móviles, … etc.

Alexa Internet, Inc. es una empresa filial de Amazon que provee las estadísticas de las visitas de un sitio web y los clasifica en un ranking. Esta información la obtiene a través de los usuarios que tienen instalada la *Alexa Toolbar*.

Se han omitido portales de anuncios clasificados que incluyen oferta inmobiliaria dentro del catálogo de servicios, por la imposibilidad de distinguir el tráfico correspondiente a tal servicio.

De la misma forma se han dejado fuera de este análisis los portales que aun siendo importantes en cuanto a oferta y tráfico, dirigen su actividad a los inmuebles situados en determinadas zonas geográficas como la comunidad de Madrid o Cataluña, y carecen de oferta en el resto del país.

Podemos diferenciar dos tipos de portales inmobiliarios que han sido tenidos en cuenta en este estudio: los portales con oferta de anuncios propia y los portales inmobiliarios agregadores, es decir, portales que carecen de oferta propia, su servicio de búsqueda dirige al usuario a portales externos que sí tienen oferta propia de inmuebles, consiguiendo de esta forma un número elevado de anuncios ofertados.

5.1. Introducción

La difícil situación que ha atravesado el mercado inmobiliario en nuestro país en los últimos años, en los que ha disminuido el número de viviendas iniciadas casi un 98 % desde 2006, año en el que alcanzó su máximo, y 2012; y el número de transacciones realizadas era un tercio de las que se realizaban al inicio de la crisis económica, ha reestructurado la oferta de empresas dedicadas al sector inmobiliario, reduciendo drásticamente el número de empresas dedicadas a la construcción. No obstante, el número de empresas dedicadas a la compra-venta y alquiler de inmuebles ha aumentado un 10 %, de 2008 a 2013, debido al stock de inmuebles generado por esta reducción tan drástica del número de transacciones realizadas y al stock previo al período de crisis. Este incremento se ha moderado fijándose en 2016 en un 7%.

La compra - venta y el alquiler de inmuebles desde internet ha aumentado exponencialmente debido a la gran oferta a la que puede acceder el usuario desde su casa. Una extensa descripción del inmueble y una gran cantidad de fotografías del mismo pueden hacer que el potencial comprador tenga una idea clara del inmueble que está "visitando". Es claro, que, en la segunda parte de la operación, y una vez que el usuario conoce virtualmente el inmueble, se produce la visita in situ del mismo.

En este contexto de crisis económica, ha cambiado en gran medida, también, el mapa de portales inmobiliarios en la red. Portales de gran importancia han desaparecido o han sido absorbidos por otros, y a su vez, nuevos portales inmobiliarios se han situado en posiciones importantes en el mercado. También se han creado numerosos acuerdos entre empresas propietarias de portales inmobiliarios para compartir sus bases de datos de anuncios.

5.2. Portales inmobiliarios con oferta propia

En primer lugar, analizaremos los portales inmobiliarios con oferta propia más relevantes actualmente en España por oferta de inmuebles, tráfico web y por representatividad dentro del mercado. No obstante, en ocasiones esta oferta es compartida con otros portales con los que tiene convenios.

El estudio de campo se realizó entre el 21 y el 29 de marzo de 2015, y la selección considerada, presentada a continuación, fue realizada en función de la importancia de los portales web inmobiliarios. De modo que, los portales analizados finalmente fueron idealista.com, fotocasa.es, hogaria.net, pisos.com, yaencontre.com, globaliza.com, expocasa.com y urbaniza.com.

5.2.1. Idealista.com

Es un portal inmobiliario que tiene como principales accionistas a Kutxabank, S.A. (15 %) y Bonsai Venture Capital (11 %). Fue creada por Jesús Encinar en el año 2000 y desde sus inicios siempre ha mantenido una línea exclusivamente inmobiliaria. La empresa propietaria del portal es Idealista, Libertad y Control, S.A. con CIF A-82505660.

En 2015, Idealista.com cerró la compra del 35% de www.tercerob.com la web que recopila la información pública del sector inmobiliario de multitud de bases de datos relevantes como la administración y los servicios haciéndola accesible para todos los usuarios.

Una empresa participada por Idealista, Libertad y Control, S.A. es www.rentalia.es, un importante portal de anuncios de alquiler vacacional.

Como puede observarse en la *Figura 32*, la web es visualmente agradable, con riqueza de imágenes, y de fácil manejo. El usuario puede escoger entre los tipos de anuncios: comprar, alquilar y compartir. Los tipos de inmuebles son Obra nueva, viviendas, vacacional, habitación, oficinas, locales o naves, garajes o terrenos. El usuario puede elegir entre todas las provincias españolas, zonas del sur de Francia y Portugal. Con el uso de las cookies, el portal guarda la última búsqueda del usuario.

Figura 32. Página principal del portal idealista. Marzo de 2015. Fuente: www.idealista.com

Los vendedores, al igual que en la mayoría de los portales, disponen de hasta dos anuncios gratuitos. En este caso, el portal limita a 30 días la gratuidad de los anuncios en determinadas zonas exclusivas de Madrid.

A fecha 21 de marzo de 2015, el portal contenía un total de 671.345 viviendas a la venta en el territorio nacional. El portal está preparado para la navegación en español, catalán, inglés, francés, alemán, italiano y portugués. La adaptación a dispositivos móviles es menos atractiva que la web original, pero es funcional y con manejo fluido. A finales de junio de 2017, el número de viviendas a la venta ascendía ya a 809.697, lo que supone un considerable incremento de más del 20% en veintisiete meses, que, unido a otros factores, afianza a este portal como la empresa líder de su sector en España.

Disponía, además, de una aplicación para dispositivos móviles, que, para los dispositivos con sistema Android cuenta ya con más de 500.000 descargas y una valoración media de 3,8 para un total de 4.261 comentarios. Se encontraba en un nivel de madurez bajo y los comentarios negativos se centraban en problemas de funcionamiento y en la falta de opciones en las búsquedas.

Este portal dispone de una aplicación API para el acceso desde el exterior a sus bases de datos, para desarrolladores. A través de este enlace: http://www.idealista.com/labs/api.htm se puede solicitar una clave de acceso a su base de datos. Con algunos conocimientos de programación, se puede desarrollar una aplicación de recogida de datos de sus inmuebles, previa autorización de la empresa.

Idealista.com estaba situado en la posición 2001 del ranking mundial Alexa con un total de 510.917 visitantes únicos diarios y una valoración según este tráfico de 4.413.960 $.

5.2.2. Fotocasa.es

Es el portal inmobiliario del grupo *Anuntis Segunda Mano*, propiedad de Tradder Media, una multinacional que fue adquirida por el grupo noruego Schibsted, propietario a su vez del portal Compraventa.com y del periódico gratuito '20 minutos'.

La descripción contenida en su web es:

fotocasa.es es el principal portal inmobiliario de España especializado en la compraventa y alquiler de viviendas de segunda mano y de obra nueva. El portal cuenta con una oferta de más de 1.000.000 de anuncios y más de 5 millones de usuarios únicos al mes. Actualmente, fotocasa.es pertenece a Schibsted Classified Media, grupo líder en España con portales como Segundamano.es, Infojobs o Coches.net.

Figura 33. Página principal del portal Fotocasa. Marzo de 2015. Fuente: www.fotocasa.es

En sus más de 15 años de trayectoria, fotocasa.es se ha consolidado como una de las webs inmobiliarias de referencia en Internet a la hora de buscar vivienda y en la actualidad es el portal inmobiliario más visitado de España, según datos de AT Internet.

Es un portal intuitivo, fácil de manejar con una barra de búsqueda en el inicio que ocupa toda la página. Está disponible para dispositivos móviles y en español, catalán, inglés y alemán. Permite seleccionar entre Comprar, alquilar, compartir y vacacional (para alquileres en estos períodos).

La búsqueda puede realizarse escribiendo la zona geográfica deseada, en la barra de búsqueda mostrada en la *Figura 33* directamente o a través de selección en un mapa. A fecha 21 de marzo de 2015, el número de viviendas en venta es de 522.432. La revisión de este dato a finales de junio de 2017 revela que este número ha incrementado desde entonces hasta los 600.041, lo que supone algo menos de un 15% en un período de nueve trimestres.

Este portal puede ser usado para la promoción de inmuebles tanto por parte de particulares como por profesionales. Permite la selección de diferentes tipos de inmuebles como pisos, apartamentos, áticos, dúplex, estudios, lofts, casas-chalet, plantas bajas, fincas rústicas, casas adosadas, oficinas, garajes, locales comerciales y terrenos.

Los particulares pueden publicar hasta dos anuncios de forma gratuita. Cuenta con un espacio especializado en hipotecas donde ofrece información actualizada sobre los productos hipotecarios, así como calculadoras. Cuenta también con un blog con información sobre el sector con noticias sobre alquiler, venta, hipotecas, reformas, suministros y decoración, donde se ofrecen ideas para el futuro hogar

La empresa realiza controles de calidad para evitar la publicidad de ofertas fraudulentas o falsas. Entre otros, realiza llamadas telefónicas a los anunciantes para comprobar que el teléfono de contacto es adecuado. Ningún anuncio es publicado en el portal si no ha pasado previamente todos estos controles, tras lo cual el anunciante recibe un código de validación que permite la visualización del anuncio.

Dispone, en la fecha de realización del trabajo, de una aplicación para dispositivos móviles. En la tienda virtual Play Store de Android acumula más de 1 millón de descargas. Es una de las aplicaciones de consulta de inmuebles más valorada y la más descargada. Su valoración para estos dispositivos es de 4 sobre 5 para un total de 20.939 opiniones. Todavía presentaba, entonces, graves problemas de funcionamiento ocasionales, ya que su nivel de madurez es bajo.

Desde enero de 2005, fotocasa.es elabora el Índice Inmobiliario fotocasa.es que tiene como objeto medir la evolución del precio de venta de la vivienda de segunda mano en España. Para ello, en febrero de 2005, la empresa estableció un valor inicial de 1000. A partir de este valor, y con la información de la base de datos de los inmuebles del propio portal, se obtienen las variaciones en el precio de la vivienda con periodicidad mensual. El último índice publicado en mayo de 2015, es de 681, lo que significa un precio medio de la vivienda de 1.635 €/m^2. Este índice se ha convertido en fuente de referencia para el Fondo Monetario Internacional (FMI) en estudios sobre la economía española.

El portal fotocasa.es estaba situado en el puesto 3.389 del ranking Alexa que mide el número de visitantes de este, y tenía un total de 301.666 visitantes únicos diarios y una valoración estimada por su tráfico de 2.606.040 $.

5.2.3. Hogaria.net

Hogaria.net es un portal inmobiliario propietario de la empresa *World Enred, S.L.*, con sede social en Madrid, que incluye anuncios de casas en venta y en alquiler y alojamientos rurales en toda España. Nace en el año 2003 con capital privado. Su actual gerente es D. César Piñeiro Álvarez.

Ofrece la posibilidad de publicación de anuncios a particulares y profesionales. Los primeros pueden publicar los anuncios por un tiempo ilimitado, pero al principio del proceso de registro del anuncio deben enviar un SMS con un coste de algo más de 1 €. Los clientes profesionales pagan una tarifa plana que les incluye servicios como una aplicación de gestión o consultoría inmobiliaria y publicidad.

El portal dispone, además, de un indicador inmobiliario que mide el precio en euros por metro cuadrado de las viviendas de segunda mano tanto a nivel nacional como provincial, con periodicidad mensual. Este indicador refleja el precio medio por metro cuadrado de las viviendas anunciadas en su web, descartando las viviendas con superficie inferior a los 50 m^2 y superior a los 200 m^2. También son eliminados para el cálculo los inmuebles con precio de venta inferior a 50.000 € y superior a 500.000 €. Se elabora un indicador nacional y también indicadores por provincias y poblaciones siempre que en ésta la muestra de inmuebles disponible por el portal sea de al menos 80 anuncios. Este

indicador goza de credibilidad a tenor de las menciones que realizan otras publicaciones web. Agencias de comunicación como Europapress la consideran una fuente de información fiable. Portales de información económica como Expansión y el economista reflejan a menudo los resultados arrojados por este indicador.

Figura 34. Página principal del portal Hogaria. Marzo de 2015. Fuente: www.hogaria.net

La apariencia del portal, como se observa en la *Figura 34*, está algo desfasada respecto a otros portales de la competencia. El buscador es, como se puede observar, el más usado por el resto de portales, y en él se puede escoger entre diferentes operaciones: Venta, alquiler, alquiler opción compra, traspaso y compartir. En tipo de inmueble las opciones son obra nueva, viviendas 2ª mano, locales/oficinas/naves, edificios/negocios y garajes/otros. Una vez seleccionadas estas opciones, marcaremos la provincia sobre la que se desea realizar la búsqueda.

También nos da la opción de realizar la búsqueda por palabras de forma libre o incluso realizar la búsqueda por referencia o teléfono del anunciante.

En la fecha en la que se realizó el trabajo de campo, - 29 de marzo de 2015 - disponía de un total de 502.012 anuncios de viviendas a la venta. Aunque la desfasada apariencia del portal, comentada anteriormente, fue actualizada meses después de finalizar el trabajo de campo, revisada la información relativa al número de viviendas a la venta, 27 meses después, este dato se había desplomado hasta las 59.952 viviendas, representando estas menos del 12 % de las medidas en el año 2015.

No disponía de aplicación para dispositivos móviles.

En cuanto a número de visitas, este portal estaba en profunda recesión en los meses previos al estudio, por lo que ocupaba el puesto 597.138 del ranking Alexa, con un número de visitantes únicos diarios de 806 y una valoración estimada por este tráfico de 1.200 $.

5.2.4. Pisos.com

Es el portal inmobiliario del grupo español de comunicación multimedia *Vocento*. Es el sucesor del portal sacacasa.com, El grupo Vocento compró el dominio de pisos.com por 250.000 euros y fue trasladado a dicho dominio. Este nuevo portal entró en funcionamiento el 19 de enero de 2009, y es propiedad de la empresa HabitaSoft, S.L., afincada en Granollers. Su director es Miguel Ángel Alemany, anterior Director General de Fotocasa.es.

Además de este portal inmobiliario, la empresa HabitaSoft, S.L. dispone del portal Habitat24.com, dirigido a profesionales, y un software de gestión inmobiliaria dirigido también a profesionales del sector, desde el que poder publicar directamente en su portal, pero también en los principales portales del mercado web como Idealista, Enalquiler, Yaencontre, Habitat24, Globaliza, Ventadepisos, Masprofesional y otros.

La descripción que pisos.com ofrece en su web en el momento de la realización del estudio es la siguiente:

pisos.com es un portal inmobiliario de referencia que, desde 2009 ayuda a usuarios y profesionales a vender, alquilar o buscar su hogar. Los inmuebles publicados en nuestro portal cuentan con todo tipo de detalles y además te ofrecemos todo lo necesario para que buscar, alquilar o vender tu vivienda sea fácil, rápido y eficaz. También contamos con diferentes servicios prácticos y útiles que contribuyen a facilitarte el proceso de compraventa. Además de toda la información del sector para ponerte al día de las últimas noticias inmobiliarias.

Figura 35. Página principal del portal pisos.com. Marzo de 2015. Fuente: www.pisos.com

La estética de la web es muy agradable, *Figura 35*, con un buscador muy sencillo de utilizar y muy intuitivo, pero una gran similitud al portal idealista.com visto anteriormente.

El usuario puede elegir entre comprar (de segunda mano) y alquilar. La compra de obra nueva se muestra como otra opción independiente. Los tipos de inmuebles a escoger son casas y pisos, locales y oficinas, naves, terrenos y garajes y trasteros. La búsqueda además se realiza por provincias, y una vez realizada la búsqueda permite realizar filtros sobre la búsqueda.

A fecha de realización del trabajo de campo, el 21 de marzo de 2015, en el portal se encontraban alojados un total de 467.872 anuncios de viviendas a la venta. Los usuarios pueden publicar anuncios de forma gratuita sin ninguna limitación. Veintisiete meses después, esta cifra había disminuido ligeramente hasta las 460.503 viviendas a la venta.

Dispone además de calculadora de hipoteca, información sobre los índices de referencia hipotecarios, información sobre el mercado inmobiliario, etc.

Presta, también, servicio a través de una aplicación para dispositivos móviles que tenía, a fecha de realización del trabajo, aún un nivel de maduración bajo, pero que ya había sido descargada por más de 100.000 usuarios. La valoración de estos para un total de 6.355 opiniones es de 4,2 para los dispositivos Android.

Aún con un número inferior de viviendas a la venta que el portal hogaria.net, el número de visitantes actual es muy superior, y había aumentado considerablemente en los meses previos a la realización del trabajo de campo, hasta situarse en el puesto 9.633 del ranking Alexa, con 106.129 visitantes únicos al día y una valoración de916.920 $.

5.2.5. Yaencontre.com

Portal inmobiliario fundado en el año 2000 con el objetivo de ofrecer servicio en Cataluña. Su dominio original es jahetrobat.com y la empresa propietaria es *Yaencontre – Jahetrobat, S.L.* situada en Olèrdola (Barcelona). Desde 2008 tiene la participación del grupo Godó, editor del periódico La Vanguardia. No sólo se ha centrado en el mercado de los anuncios inmobiliarios, en el año 2009 amplía su campo de acción hacia el sector de los regalos con la web Siibil.com, además de probar suerte en el mundo de los anuncios de empleo con Clasificados.es.

Figura 36. Página principal del portal yaencontre.com. Marzo de 2015. Fuente: www.yaencontre.com

El interfaz gráfico, mostrado en la *Figura 36*, es agradable, aunque la barra de búsqueda tiene menos protagonismo en la pantalla principal que en el resto de portales. Además, la carga de las páginas es más lenta de lo habitual y la búsqueda tarda más que las de otros portales analizados.

En tipo de anuncio nos permite elegir entre comprar, alquilar, compartir, alquiler con opción de compra, alquiler de temporada, traspaso y permuta. Los tipos de inmuebles a

elegir son viviendas, terrenos, oficinas, locales, naves, edificios y parkings. En lugar de seleccionar el lugar geográfico en un listado, el usuario debe escribirlo en una barra de búsqueda.

Los particulares pueden publicar hasta 2 anuncios de forma gratuita, incluso en varios idiomas distintos.

A fecha 29 de marzo de 2015 había disponibles un total de 299.924 viviendas a la venta. Dos años y tres meses más tarde, esa cifra se había reducido casi hasta la tercera parte, 119.985.

Dispone de una aplicación para dispositivos móviles que no se actualizaba desde hacía más de 5 meses, con más de 10.000 descargas realizadas para dispositivos Android, y con una valoración de 3,2 sobre 5 de los usuarios que se la habían descargado para un total de 230 comentarios.

Es un portal que también ha ganado importancia en la web en los últimos tiempos. En la fecha del estudio se encontraba en el puesto 15.911 del ranking Alexa, y a él accedían diariamente 60.474 visitantes únicos. Su valoración estimada debido a este tráfico era de 522.720 $.

5.2.6. Globaliza.com

El titular de este portal es Globaliza, S.L., empresa afincada en El Plantío, Madrid. Es el resultado de la fusión de uno de los portales inmobiliarios españoles más veteranos, fundado en 1998 como Global Habitat por Gonzalo Ortiz y Gonzalo del Pozo, y Suvivienda.es, portal inmobiliario del grupo Unidad Editorial. En septiembre de 2008 firmó un acuerdo con Yaencontre.com para compartir sus bases de datos de anuncios inmobiliarios. Esta fusión tuvo lugar en octubre de 2009.

Figura 37. Página principal del portal Globaliza. Marzo de 2015. Fuente: www.globaliza.com

La página principal del portal, *Figura 37*, consta de un buscador que coincide con el de otros buscadores ya vistos. Además, nos da la opción de dirigirnos directamente a la búsqueda de pisos de bancos o de alquileres vacacionales.

Permite la inserción, en la fecha del estudio, de anuncios por parte de particulares y de profesionales. Los anuncios se pueden insertar de forma gratuita, aunque los clientes de alquileres deben enviar un SMS con un coste de 1,43 € para dar de alta el anuncio. Además, hay un plan de pagos diseñado para priorizar el anuncio en las búsquedas.

Disponía entonces de un total de 104.204 viviendas a la venta y 59.078 pisos de bancos. Este primer número se había incrementado ligeramente hasta las 119.433 viviendas a la venta en junio de 2017.

Se estimaba, entonces, que el número de visitantes únicos diarios del portal era de 19.800 alcanzando con esto el puesto 48.596 del ranking mundial Alexa, con una valoración de 171.360 $.

5.2.7. Expocasa.com

Es uno de los portales inmobiliarios más antiguos del mercado. Pertenece al grupo *Facilísimo Interactive, S. L.* y fue fundado en 1999 por Alberto Fernández, que actualmente conserva la mayoría de las participaciones del grupo.

Entre 2008 y 2013 se integró como un canal más dentro de los verticales de facilisimo.com, y en septiembre de 2013 recupera su independencia como portal, aunque continúa como parte del Grupo Facilísimo.

Facilisimo.com es una compañía española centrada exclusivamente en internet que está formada por un conjunto de canales temáticos dedicados al hogar (decoración, cocina, belleza, jardinería, inmobiliaria etc.). Además, cada uno de sus canales temáticos incorpora un área de comunidad donde los usuarios pueden intercambiar ideas y opiniones.

Figura 38. Página principal del portal expocasa.com. Marzo de 2015. Fuente: www.expocasa.com

La *Figura 38* muestra una página sencilla y fácil de utilizar. La búsqueda se realiza por tipo de inmueble (viviendas, garajes, oficinas, naves, locales y terrenos), por tipo de anuncio (en venta, en alquiler y a estrenar) e insertando la provincia o el código postal. También se puede realizar la búsqueda por teléfono del anunciante o por número de referencia. Los vendedores particulares pueden insertar 2 anuncios de forma gratuita.

No disponía de aplicación para dispositivos móviles.

A fecha 21 de marzo había publicados un total de 138.273 anuncios de venta de viviendas, con un precio medio de 1519 € / m^2. Esta oferta ha ido incrementándose hasta alcanzar la cifra de 312.928 viviendas. Sólo tiene opción de idioma español y la adaptación del portal a dispositivos móviles presenta muchas carencias.

Es un portal que, respecto al número de visitantes, ha ido disminuyendo paulatinamente de importancia en el sector. No en vano, ocupa el puesto 245.308 del ranking Alexa y tiene sólo 3.432 visitantes únicos diarios y una valoración de 20.520 $. Sin embargo, la oferta inmobiliaria continúa creciendo.

5.2.8. Urbaniza.com

Este portal fue creado en mayo de 2000 y pertenece al grupo inversor del ya desaparecido portal Guay.com. En agosto de ese mismo año se crea la empresa Urbaniza Interactiva, S.A. con una actividad principal que es la gestión publicitaria en internet, pero con otras actividades como desarrollos en internet, sistemas de gestión, creación y desarrollo de webs inmobiliarias, provisión de contenidos y consultoría en negocios online para el mercado inmobiliario. En septiembre de 2001 comenzó a ser el canal inmobiliario de varias cajas de ahorros, como Caja Duero, Caja de Ávila o Caja Cantabria.

El portal, a través de acuerdos de colaboración con otros portales, que le han permitido compartir bases de datos de inmuebles, ha crecido de manera que actualmente cuenta con una gran bolsa inmobiliaria.

En junio 2002 firma un acuerdo con CECA, por el que se designa a Urbaniza como proveedor homologado para la creación y gestión de portales inmobiliarios para entidades financieras y poco después lanza el primer portal para una entidad financiera: CajaCantabria.

A principios de 2005 ya cuenta con canales inmobiliarios para CAN, CajaDuero, Cajasur, Caja El Monte y CajaCanarias y CajaMadrid. Durante 2009 lanza las webs de Altamira Real Estate, activos adjudicados Santander y Banesto Vivienda, y el portal inmobiliario del Banco Cooperativo y Cajas Rurales.

Durante 2011 y 2012 lanza el portal web y el sistema de gestión web de bbvavivienda.com, rediseña relanza el portal inmobiliario de NovaGalicia: escogecasa.es y pone en marcha el portal inmobiliario de Banco Espirito Santo.

Figura 39. Página principal del portal Urbaniza. Marzo de 2015. (Fuente: www.urbaniza.com)

La web, mostrada en la *Figura 39*, nos permite buscar a través del mapa señalando la ubicación o bien eligiendo la provincia y/o la población deseada. Nos permite seleccionar el tipo de anuncio, entre venta, alquiler, traspaso y alquiler con opción de compra, así como acotar directamente la búsqueda a número mínimo de habitaciones y precio máximo. Una vez en los resultados de búsqueda, estos pueden filtrarse además por número mínimo de baños y superficie mínima.

Estéticamente es algo desfasada, lo que es un indicador de las horas bajas por las que pasa actualmente este portal. El número de viviendas en venta es bastante limitado, a fecha 29 de marzo de 2015, oferta un total de 19.481 viviendas, de las cuales 15.429 son pisos y 4.052 casas/chalets. Además, gran parte de su oferta actual se sitúa al norte de España.

Disponía, durante el trabajo de campo, de una aplicación para dispositivos móviles donde se define como un buscador inmobiliario similar a idealista o fotocasa, pero con viviendas exclusivas. No se actualiza desde abril de 2012 y se ha descargado entre 5.000 y 10.000 veces. La valoración, para 28 usuarios, es de 2,1 sobre 5. La mayoría de las valoraciones es de 1 (mínima) y los comentarios muy negativos.

Es un portal de segundo nivel y no es comparable con los otros portales inmobiliarios considerados anteriormente, no obstante, se ha incluido en este estudio debido a la fuerte relación de la empresa propietaria con los portales comercializadores de activos inmobiliarios procedentes de entidades bancarias. De hecho, según el ranking Alexa, este

portal está valorado en tan sólo 240 $, ya que tiene un tráfico de 224 visitantes únicos al día y ocupa el puesto 2.133.237 a nivel mundial.

5.3. Comparativa entre los portales con oferta propia analizados

Realizaremos a continuación una comparativa entre los portales analizados con el objetivo de crear una clasificación de los portales en función de su importancia. Debido a que cada portal oferta diferentes tipos de anuncios, se ha elegido como medida de comparación de la oferta de anuncios de vivienda el más representativo de cada uno, la venta de viviendas de segunda mano. Estas cifras, recogidas a finales del mes de marzo de 2015, fluctúan constantemente, por lo que se han tomado en un período corto de tiempo. Son las que se muestran en la Tabla *17*:

Tabla 17.*Comparativa portales. Oferta de viviendas a la venta. Marzo 2015*

Portal	**N.º viviendas a la venta Marzo de 2015**	**N.º viviendas a la venta Junio de 2017**
Idealista.com	671.345	809.697
Fotocasa.es	522.432	600.041
Hogaria.net	502.012	59.952
Pisos.com	467.872	460.503
Yaencontre.com	299.924	119.985
Globaliza.com	163.282	119.433
Expocasa.com	138.273	312.928
Urbaniza.com	19.481	-

Fuente: Elaboración propia a partir de la información incluida en los portales

Podemos observar en la *Figura 40*, que hay 4 portales que destacan sobre los demás, que son idealista.com, que a su vez tiene casi 150.000 anuncios que el siguiente, fotocasa.es, hogaria.net y pisos.com. Entre los 4 representan el 77,7 % de la oferta total.

Aunque los dos primeros han experimentado un crecimiento en su oferta, la del portal hogaria.net se ha desplomado y la de pisos.com permanece sin grandes variaciones.

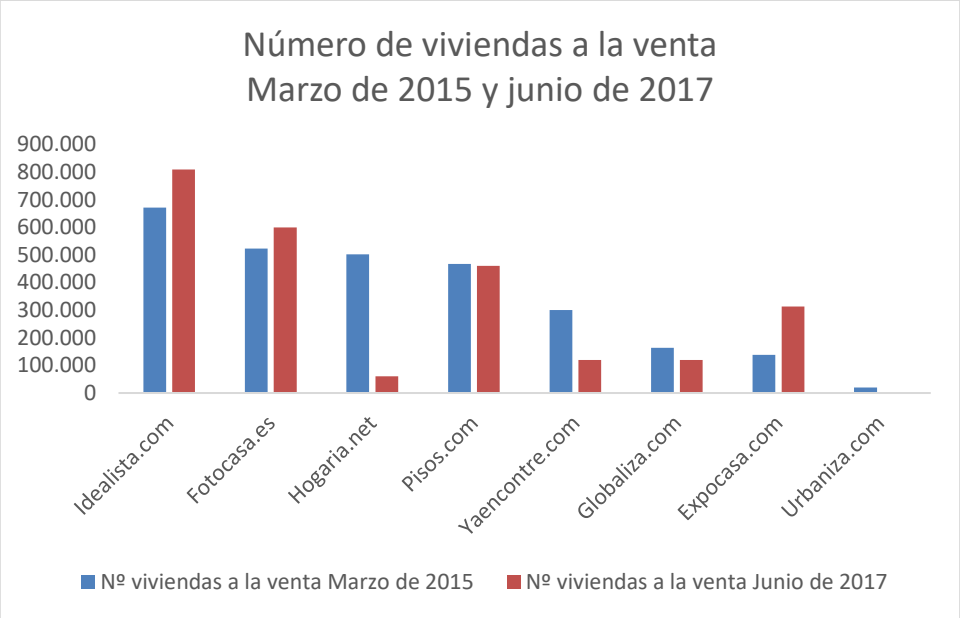


Figura 40. Comparativa portales. Oferta de viviendas a la venta. Marzo 2015

El otro indicador que consideraremos para medir la calidad de la web es el del número medio de visitantes únicos al día de la web. Estos datos se han recogido de la web sobreinternet.com que a su vez recopila la información de la web alexa.com, web operado por Alexa Internet, Inc. que es una empresa subsidiaria de Amazon. Esta empresa realiza las estadísticas de tráfico a través de la información recopilada de los usuarios que tienen instalada su barra de herramientas: Alexa Toolbar. Veamos los resultados obtenidos en la Tabla *18*:

Tabla 18. *Comparativa portales. Número de visitas únicas diarias. Marzo 2015.*

Portal	**N.º visitas únicas / día**	**Valoración ($)**
Idealista.com	510917	4.413.960
Fotocasa.es	301666	2.606.040
Hogaria.net	806	1.200
Pisos.com	106129	916.920
Yaencontre.com	60474	522.720
Globaliza.com	19800	171.360
Expocasa.com	3432	20.520
Urbaniza.com	224	240

Fuente: Elaboración propia a partir de la información recogida por Alexa

En este factor se destaca claramente, respecto a sus competidores, el portal idealista.com, que tiene más visitantes únicos diarios que el resto de los portales analizados juntos. A cierta distancia le siguen fotocasa.es y pisos.com. Entre estos 3 portales se reparte más del 90 % del total de visitas analizadas. También es destacable que el portal idealista.com tiene un valor económico superior a la suma de todos los portales con oferta propia restantes analizados.

En la *Figura 41* destaca negativamente el portal hogaria.net con el segundo dato más bajo del grupo analizado y tan sólo 806 visitantes únicos diarios con más de 500.000 viviendas a la venta.

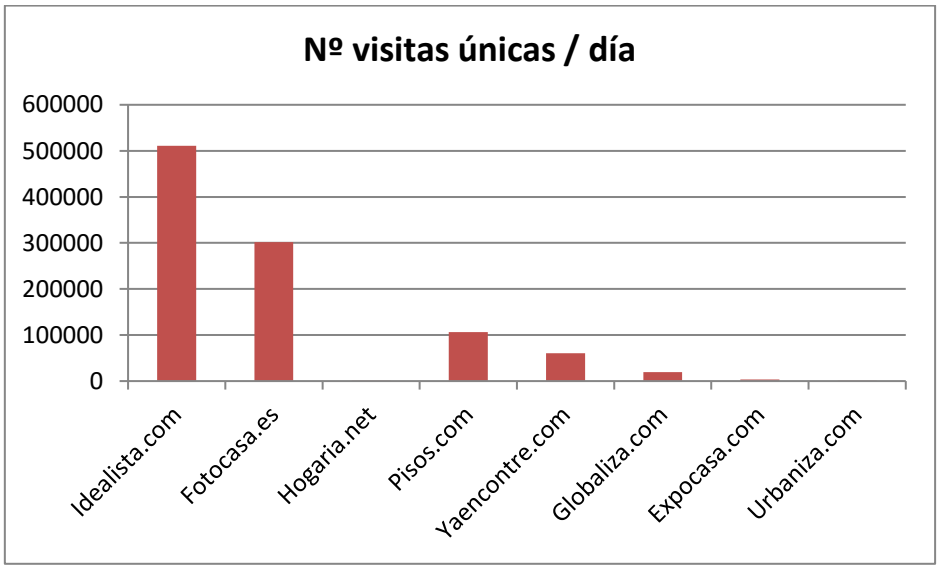


Figura 41. Comparativa portales. Número de visitas únicas diarias. Marzo 2015. Fuente: Elaboración propia a partir de los datos proporcionados por alexa.com

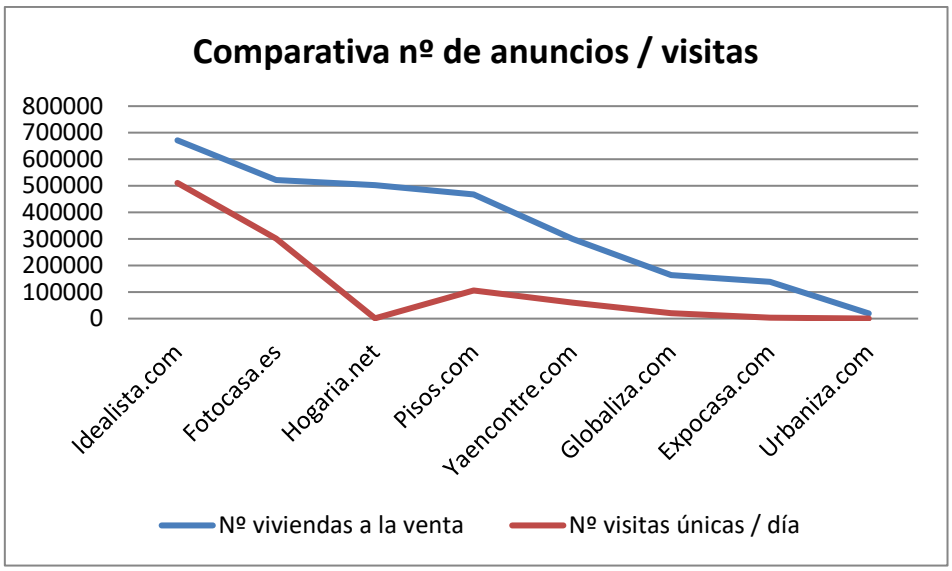


Figura 42. Comparativa portales. N.º de visitas vs. N.º de anuncios. Marzo 2015.

Por último, pueden observarse ambos factores en un mismo gráfico, como muestra la Figura 42.

Además, en la Tabla 19 se indican los principales servicios ofertados por los distintos portales.

Los portales líderes en número de visitas y en número de anuncios de venta de viviendas son también los que mejores servicios ofrecen. Prácticamente todos ofrecen la posibilidad de incluir hasta 2 anuncios por persona para particulares. Sorprende que idealista.com y fotocasa.es sean los únicos portales que tienen la web adaptada a diferentes idiomas. Respecto a la oferta de aplicaciones para dispositivos móviles, sólo idealista, pisos y fotocasa disponen de una oferta de calidad con una continua renovación.

Tabla 19. Comparativa de portales. Servicios. Marzo 2015.

Portal	Aplicación móvil	Publicación gratuita de anuncios de particulares	Multilenguaje
Idealista.com	☑ Valoración: 3,8 Descargas: + 500.000	☑ Máximo: 2	☑
Fotocasa.es	☑ Valoración: 4 Descargas: + 1.000.000	☑ Máximo: 2	☑
Hogaria.net	☒	☑ Máximo: 2	☒
Pisos.com	☑ Valoración: 4,2 Descargas: + 100.000	☑ Máximo: 2	☒
Yaencontre.com	☑ Valoración: 3,2 Descargas: + 10.000	☑ Máximo: 2	☒
Globaliza.com	☒	☑	☒
Expocasa.com	☒	☑ Máximo: 2	☒
Urbaniza.com	☑ Valoración: 2,1 Descargas: +5.000	☒	☒

Fuente: Elaboración propia a partir de la información incluida en los portales

Teniendo en cuenta todo lo anterior, podemos afirmar que el portal idealista.com es superior al resto tanto en número de anuncios de viviendas a la venta como en número de visitantes únicos diarios, y con una oferta de servicios amplia. A continuación, se sitúa fotocasa.es con una gran cartera de servicios, pero un tráfico y una oferta significativamente inferior. En tercer lugar, se sitúa pisos.com ya que, aunque hogaria.net tenía más anuncios, estos se han desplomado. Además, el tráfico web y la oferta de servicios de pisos.com son superiores.

5.4. Portales agregadores de la oferta

A continuación, vamos a analizar los principales portales agregadores de anuncios inmobiliarios. Al unificar los resultados de diferentes portales inmobiliarios, es normal que el número de anuncios sea superior. No obstante, cabe destacar que pisos.com es el portal colaborador común a todos estos en estos portales agregadores.

5.4.1. Pisos.mitula.com

Este portal es un agregador de anuncios de otros portales inmobiliarios. Forma parte de la web de anuncios mitula.com con anuncios de inmuebles, coches, trabajo y alquileres vacacionales. La empresa propietaria es *Mitula Classified, S.L.* con sede en Madrid y que inició su actividad hace casi 6 años. Su objeto social es la explotación de agregadores de portales de internet, factura en 2015 más de 3.000.000 € al año y está presente en 38 países.

Pisos.mitula.com es uno de los portales agregadores más importantes a nivel nacional. Aunque muestra anuncios de un gran número de portales, su principal *partner* es pisos.com, algo que ya hemos visto en otros agregadores.

Figura 43. Página principal del portal Mitula. Marzo de 2015. (Fuente: www.mitula.com)

Es un buscador genérico, similar, como puede apreciarse en la *Figura 43*, para todos los servicios del portal en el que la búsqueda es libre. El usuario puede escribir el lugar de la búsqueda o las características del inmueble que desea localizar. En el tipo de

anuncio, se puede elegir entre venta, alquiler y compartir. Sin embargo, está selección no es tenida en cuenta en el resultado de la búsqueda, ya que esta nos muestra los 3 tipos de anuncios. Por tanto, si deseamos sólo uno de estos tipos de anuncios, en los resultados de la búsqueda tendremos que filtrar de nuevo por tipo de anuncio.

No se permite la publicación directa de anuncios en el portal, debido a que sólo actúa como enlace. Sin embargo, sí que da la opción en la web. Al seguir el enlace, nos redirige a pisos.com.

El 29 de marzo de 2015, el portal ofrecía un total de 4.106.571 viviendas a la venta en toda España, por lo que es el portal agregador más importante en oferta. Revisiones posteriores han devuelto cifras similares.

Los partners principales son pisos.com, globaliza.com, tucasa.com, yaencontre.com… No obstante, dispone de cientos de webs asociados, es más, en la propia web hay un formulario para que cualquier partner que desee ser indexado pueda solicitarlo.

La sección inmobiliaria de este portal dispone de una aplicación para dispositivos móviles: Mitula Casas. De ella se habían realizado, en marzo de 2015, más de 50.000 descargas y se encuentra en un nivel de madurez bajo. Respecto a la valoración de los usuarios, presentaba 353 valoraciones con una media de 3.6 sobre 5.

Al pertenecer al dominio mitula.com, sólo se dispone de estadísticas de tráfico de todo el portal y no puede separarse del resto. El portal estaba en el puesto 18.753 del ranking Alexa con 51.310 visitantes únicos diarios, y una valoración de 443.520 $.

Se ha incluido en el estudio este portal, aun formando parte de un portal de anuncios genérico debido a que la principal actividad de este portal es la inmobiliaria.

5.4.2. Nuroa.es

Nuroa es un portal inmobiliario fundado en 2006, al igual que la empresa propietaria, Nuroa Internet, S. L., con sede en Barcelona. Es, al igual que el resto, un portal agregador que reúne anuncios de diversos portales inmobiliarios: pisos.com (principal colaborador),

yaencontre.com, solvia.es, idealista.com, fotocasa.es, globaliza.com, casanuncio.com - que no ha sido incluido en este estudio por ser un portal con 158.473 anuncios de venta de viviendas de las que más de la mitad se centran en la provincia de Alicante - ..., y algunos portales de bancos como servihabitat.com, así hasta un total de 66 portales.

En 2008, la firma de capital de riesgo *Highgrowth Partners* entra en Nuroa como inversor, y en mayo de 2010 inyecta en la compañía 1 millón de euros de capital que permite su rápido crecimiento.

En 2015, Nuroa cuenta con presencia en 15 países, como son Austria, Alemania, Francia, México, Brasil, Argentina, Reino Unido o Portugal. En 2011 creó una importante plataforma de alquiler vacacional www.migoa.com

Figura 44. Página principal del portal Nuroa. Marzo de 2015. (Fuente: www.nuroa.es)

La página principal es básica, con una barra de búsqueda que ocupa toda la pantalla con posibilidad de realizar una búsqueda libre por provincia, barrio o tipo de vivienda buscada (Figura 44). Se puede elegir entre venta, obra nueva, alquiler, compartir y vacaciones. También se puede ver un contador del número total de anuncios a los que enlaza este agregador.

El portal da la opción de insertar anuncios, pero al acceder a esta opción, el portal nos dirige a su principal colaborador: pisos.com asegurándonos que en 24 horas aparecerá nuestro anuncio publicado en nuroa.es.

El 29 de marzo el número de anuncios de venta de viviendas era de 3.066.092 de los cuales, 2.129.718 eran pisos y 936.374 casas. No obstante, este recuento está sobredimensionado ya que tras diferentes pruebas se comprobó que era habitual la inclusión de anuncios que no se corresponden con las búsquedas solicitadas, como garajes o trasteros; además, algunos anuncios aparecen duplicados en los diferentes partners.

No dispone de aplicación para dispositivos móviles.

Nuroa.es estaba valorado en 183.600 $ por su tráfico y ocupaba el puesto 45.237 del ranking mundial Alexa. El número de visitantes únicos diarios era de 21.270.

5.4.3. Nestoria.es

Nestoria.es es un portal perteneciente a la empresa *Lokku Limited*, afincada en Londres, Reino Unido. Es un portal agregador que busca entre los anuncios profesionales de diversos portales inmobiliarios como expocasa.com, idealista.com, pisos.com o yaencontre.com, por lo que no es posible incluir un anuncio en su web. Esto lo indican de la siguiente manera en su propia web: *No insertamos anuncios de pisos y casas de usuarios particulares. Si tienes un piso o casa que vender o alquilar te recomendamos anunciarlo en un portal inmobiliario de calidad. Será muy probable que tu vivienda aparezca en nuestro sitio web gracias a nuestros acuerdos de colaboración.*

La propia empresa define el portal como un buscador temático que cuenta con el apoyo de importantes portales inmobiliarios de toda España. Esta empresa dispone de un portal en el Reino Unido, *London and UK homes for sale*, que es referencia no sólo a nivel nacional, también a nivel europeo.

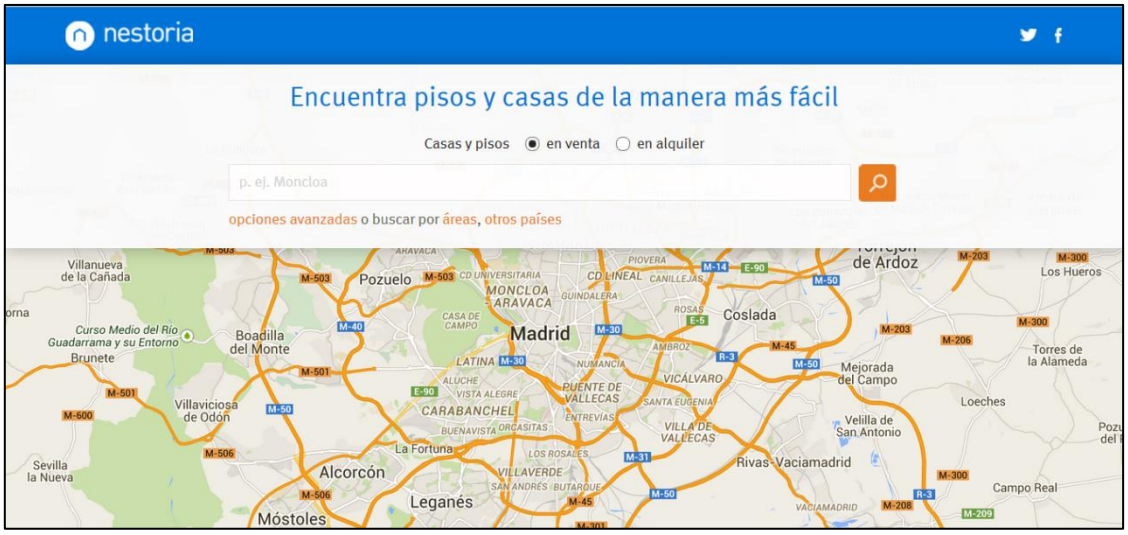


Figura 45. Página principal del portal Nestoria. Marzo de 2015. (Fuente: www.nestoria.es)

La página de inicio es sencilla con un buscador que sólo permite seleccionar entre venta y alquiler, aunque dispone de una opción de búsqueda avanzada con más opciones. Como se ve en la *Figura 45*, casi toda la página la ocupa un mapa en el que seleccionar la región de búsqueda.

Para cada resultado, el portal indica en que web se encuentra el anuncio y si se solicita más información sobre el mismo redirige automáticamente al portal en el que éste se encuentra anunciado. Es difícil identificar el número de anuncios que este portal tiene indexados debido a que cuando un inmueble está publicado en más de un lugar, el anuncio aparece en los resultados repetido. Además, aparecen, en la búsqueda, inmuebles no solicitados como, por ejemplo, terrenos rústicos. El número de resultados para compra de vivienda a 29 de marzo de 2015 era de 1.148.595.

No dispone de aplicación para dispositivos móviles.

Ocupaba el puesto 72.257 del ranking Alexa con 13.316 visitas únicas diarias. El dominio está valorado, por esto, en 115.200 $.

5.4.4. Tucasa.com

Es un portal inmobiliario que tiene su propia base de datos de anuncios, y que, además, es un portal agregador. La propietaria es la empresa *Iberanuncios S.L*. con sede en Madrid y una antigüedad de algo más de 5 años en el momento del estudio.

Es un portal que oferta sus propios anuncios, y que además enlaza con otros portales de anuncios como pisos.com, su principal partner, y a diversos portales de oferta de activos inmobiliarios de entidades bancarias como Servihabitat (empresa comercializadora de La Caixa) o Solvia (Banco Santander)

Dispone de un servicio no visto en otros portales que es el de buscar compradores, en el que en función de la zona donde se sitúe el inmueble que el cliente desea vender, se le suministra un listado de personas que han mostrado su intención de adquirir un inmueble en ese lugar.

Además, si un cliente publica un anuncio en tucasa.com, este anuncio también será publicado en otros agregadores como casas.trovit.es, nestoria.es y nuroa.es. También, en un gran número de diarios locales.

La apariencia es sencilla y dinámica, y el buscador es muy sencillo. Como puede observarse en la *Figura 46*, en un desplegable, situado en la parte izquierda de la ventana, elegiremos el tipo de inmueble a buscar: vivienda, obra nueva, oficina, local comercial, garaje, terreno, nave o trastero. A continuación, escribiremos la zona geográfica donde deseamos realizar la búsqueda.

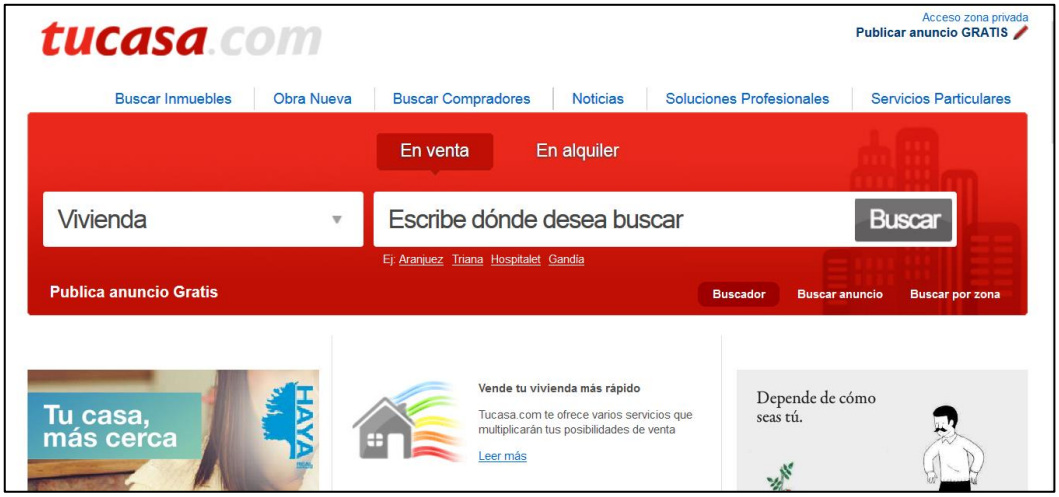


Figura 46. Página principal del portal Tucasa.com. Marzo de 2015. (Fuente: www.tucasa.com)

También podemos realizar la búsqueda de un anuncio concreto indicando el teléfono o el nombre del anunciante, así como elegir el destino entre una lista de provincias españolas.

Cualquier particular puede publicar de manera gratuita hasta un máximo de 5 anuncios de forma gratuita, sin embargo, existen diversos servicios para potenciar el anuncio que son de pago. En el momento de la elaboración del trabajo de campo, este portal tenía un total de 473.210 anuncios de viviendas a la venta.

No dispone de aplicación para dispositivos móviles.

El número de visitantes de esta página aumenta día a día, y ya estaba, entonces, en el puesto 26.802 del ranking mundial Alexa con 35901 visitantes únicos diarios y una valoración, según el tráfico, de 310.320 $.

5.5. Comparativa entre los 4 portales agregadores analizados

En primer lugar, cabe destacar que el portal tucasa.com es un portal mixto, es decir, enlaza anuncios de otras webs y tiene una base de datos propia. Los resultados para el número de anuncios enlazados se muestran en la Tabla *20*:

Tabla 20. *Comparativa de agregadores. Número de viviendas a la venta.*

Portal	**N.º viviendas a la venta Marzo de 2015**
pisos.mitula	4.106.571
Nuroa	3.066.092
Nestoria	1.148.595
tucasa	473.210

Fuente: Elaboración propia a partir de la información disponible en los portales

Los portales con mayor número de anuncios son pisos.mitula y Nuroa con más de 7 millones. Por otro lado, el portal con menor número es tucasa.com con menos de 500.000. Un diagrama de barras con esta información puede verse en la *Figura 47*.

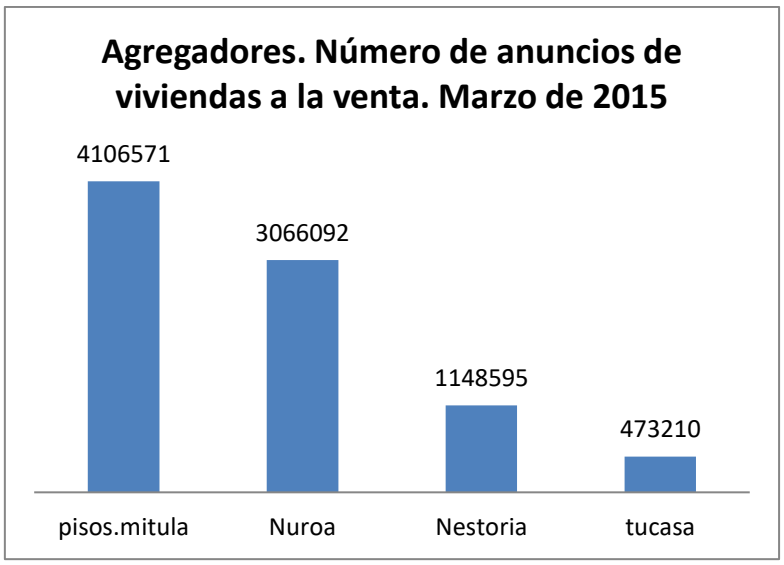


Figura 47. Comparativa de agregadores. Número de viviendas a la venta.

Al analizar el número de visitantes únicos diarios, los resultados obtenidos se muestran en la Tabla *21*:

Tabla 21.*Comparativa de agregadores. Número de visitas únicas diarias.*

Marzo de 2015		
Portal	**N.º visitas únicas / día**	**Valoración ($)**
pisos.mitula	51.310	443.520
Nuroa	21.270	183.600
Nestoria	13.316	115.200
Tucasa	35.901	310.320

Fuente: Elaboración propia a partir de los datos proporcionados por Alexa.com

El portal con más visitas únicas es mitula.com. No obstante, y como se ha indicado anteriormente, este dato corresponde al portal completo y no sólo a la sección de anuncios inmobiliarios. El siguiente portal en visitas es tucasa.com, que muestra el rápido aumento que está experimentando en el último tiempo con una oferta inmobiliaria aún pobre respecto a sus competidores. (Véase *Figura 48*)

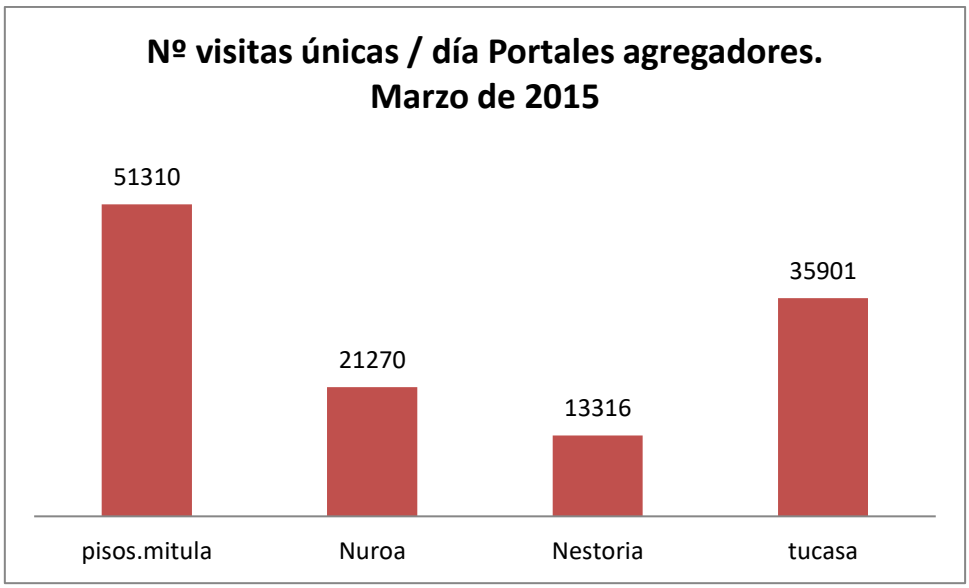


Figura 48. Comparativa de agregadores. Número de visitas únicas diarias. (Fuente: Elaboración propia a partir de los datos proporcionados por Alexa.com)

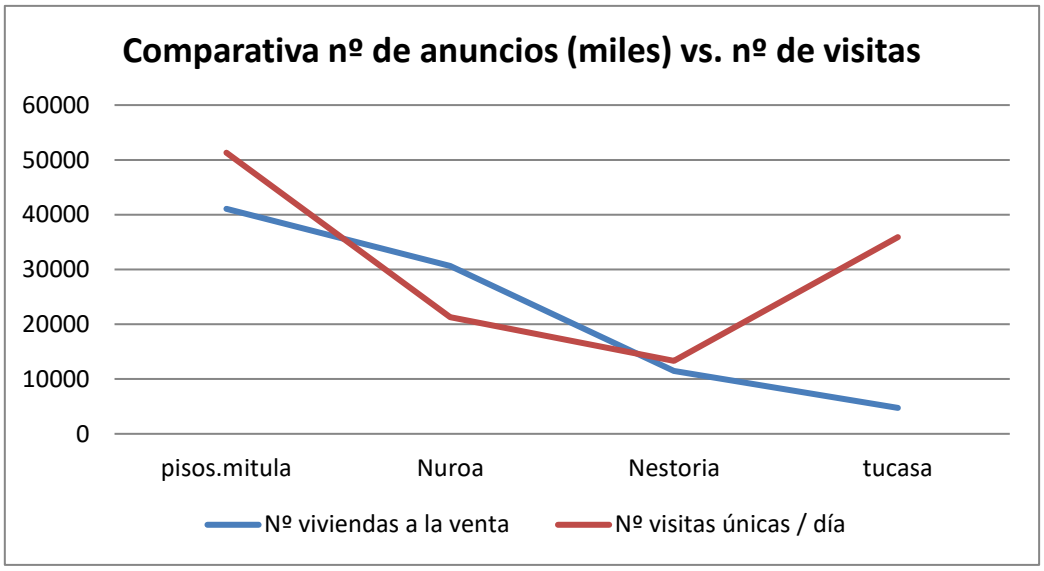


Figura 49. Comparativa agregadores. N.º de visitas vs N.º de anuncios. (Fuente: Elaboración propia a partir de los datos proporcionados por Alexa.com)

Es de suponer que al igual que en el caso de los portales inmobiliarios, un aumento de la oferta de inmuebles de un agregador, conllevará un aumento del número de visitas. Para ello, mostramos en la *Figura 49* el resultado de ambas variables en un único gráfico, en el que se han modificado las escalas para poder unir estos resultados.

Observamos que en general un mayor número de viviendas a la venta supone un mayor número de visitantes únicos diarios, sin embargo, en el agregador tucasa.es con un

número significativamente inferior de anuncios, tiene una cifra de visitas superior por ejemplo a nuroa.es con más de 3 millones de anuncios de viviendas a la venta.

Respecto a los servicios de los portales agregadores cabe destacar que no es posible publicar anuncios en los mismos debido a que no disponen de una oferta propia. En algunos de ellos se redirige al usuario a los portales inmobiliarios proveedores.

En ninguno de los portales agregadores se puede seleccionar un idioma distinto al castellano, y sólo pisos.mitula.com dispone de una aplicación para dispositivos web con un alcance medio y un nivel de desarrollo bajo.

6.El programa

Este programa tiene como objetivo proporcionar una herramienta integral para el análisis estadístico del mercado inmobiliario. Para ello, proporciona un sistema de recogida de información de inmuebles en oferta del portal web líder en el sector en España, un sistema de filtrado y exportación de datos y, por último, una completa herramienta de análisis estadístico basada en el motor del software de libre distribución R.

En una primera parte, analizaremos la estructura del programa, detallando su funcionamiento desde el punto de vista técnico y especificando el elaborado proceso de búsqueda que se ha implementado para solventar las limitaciones existentes en los servidores que contienen los datos. En la segunda parte, analizaremos en detalle los tipos de proyectos que pueden llevarse a cabo y las variables recogidas en cada uno de ellos. En la tercera parte, se hace un recorrido por el proceso de instalación del programa[2] a modo de guía de usuario. Por último, se efectúa una explicación de cada una de las

[2] Todas las imágenes mostradas son capturas del programa y por tanto son de elaboración propia.

ventanas de que se compone el programa para que sirva de guía de funcionamiento del mismo.

6.1. Detalles técnicos de funcionamiento del software

InmoDataAnalizador es una aplicación que utiliza diversas tecnologías de programación. Su núcleo principal está desarrollado en lenguaje de programación Java, lo que le confiere la posibilidad de utilizar la aplicación en cualquier sistema operativo compatible con él, ya que es un lenguaje multiplataforma. La parte dedicada a los estudios estadísticos está desarrollada a través de funciones implementadas en lenguaje R, lo que le otorga una gran potencia de cálculo.

Para la interfaz gráfica del programa se ha decidido utilizar JavaFX® para darle un aspecto más moderno y adaptarlo a la última tecnología para el desarrollo de interfaces en Java®. Concretamente, se ha utilizado la versión de Java 1.8.0_40, lo cual lo convierte en un requisito del sistema. (Oracle And/Or Its Affiliates, 2015). Sólo se podrá usar la aplicación en aquellos sistemas que tengan una versión igual o superior a la anteriormente citada, instalada en el equipo.

Veamos en primer lugar, un listado de los requisitos mínimos recomendados para el uso del programa, y a continuación, especificaremos su arquitectura.

6.2. Requisitos del sistema

La aplicación requiere un uso intensivo de datos para los estudios estadísticos, por ello se aconsejan los siguientes requisitos:

- Requisitos hardware.
  - Procesador de 2 GHz o superior
  - Memoria RAM de 4GB o superior
  - Conexión a internet, para el acceso a la información
- Requisitos software.
  - Sistema operativo compatible con el software R, puede ser *Windows 7* o superior, o cualquier sistema que disponga de alguna versión de R y compatible con Java 1.8.0_40 o superior.
  - Java 1.8.0_40 o superior.
  - R 3.2 o superior con la librería *rJava* instalada.

- Algún programa ofimático o similar que tenga soporte para la apertura de ficheros CSV, como *Microsoft Excel* o *OpenOffice Calc*.

6.3. Arquitectura y análisis

La aplicación está compuesta, principalmente, por tres sistemas: el sistema de adquisición de datos y almacenamiento, la interfaz de usuario y el sistema de análisis de información, entre los cuales existe una comunicación que permite el uso y manejo de la base de datos y de la información de los inmuebles contenidos en ella, tal y como puede verse en la *Figura 50*.

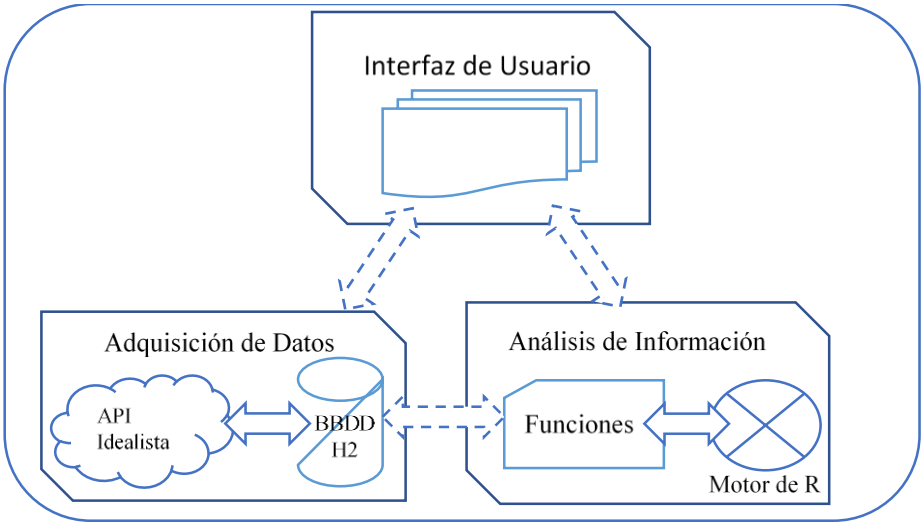


Figura 50. Sistemas que conforman la aplicación y sus relaciones.

Veamos a continuación, en detalle, cada uno de los sistemas de los que está compuesto el programa.

6.3.1. Sistema de Adquisición de datos

El sistema de adquisición de datos es un sistema con una arquitectura en tres capas: la aplicación, - que es la encargada del procesamiento de la información - la base de datos, - encargada del almacenamiento de los inmuebles - y la API (interfaz de programación de aplicaciones) de Idealista – que es el conjunto de funciones dadas por la empresa para la obtención de inmuebles desde su servidor, y que contiene una capa de abstracción -. Puede verse la interacción entre las tres capas en la *Figura 51*.

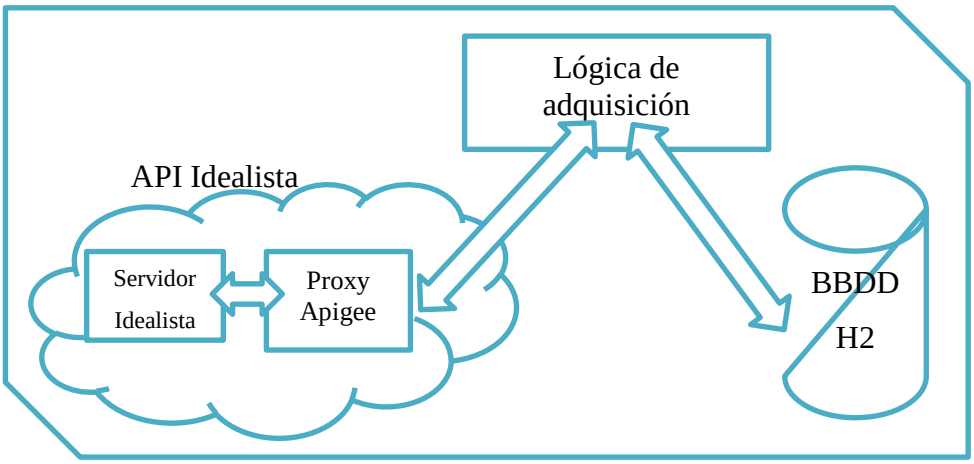


Figura 51. Capas del sistema de adquisición de datos. (Fuente: Elaboración propia)

El funcionamiento de cada una de las capas que conforman el sistema de adquisición, se muestra a continuación.

6.3.1.1. La API

El acceso a la información de los inmuebles, se hace a través de una API implementada por los desarrolladores de Idealista y que han puesto a disposición de todo aquel que quiera obtener información de su sitio web, previo permiso del administrador, a través de una clave proporcionada por el mismo.

El público destinatario de esta API, es, fundamentalmente, aquellos que se dedican profesionalmente al mundo inmobiliario, es decir, agentes de la promoción inmobiliaria que cuentan con una cartera de inmuebles que desean promocionar. El acceso a su API es un servicio de pago. No obstante, la empresa pone a la disposición de desarrolladores que deseen crear proyectos en los que se haga uso de ella, previa presentación del mismo y autorización de Idealista, como es nuestro caso. Sea como fuere, Idealista impone limitaciones en el número de peticiones que se pueden hacer al servidor, entendibles, para evitar la saturación en su sistema, que en ocasiones ha supuesto problemas en la elaboración de este trabajo. La comunicación con la empresa ha sido continua, y la ayuda reciba, desinteresada.

Debido a estas limitaciones, el sistema debe almacenar la información en una base de datos, de forma que se evite consultar desde la API reiteradamente. Ésta impone dos tipos de limitaciones importantes: una a nivel de la velocidad de las peticiones al servidor (sólo puede realizarse, a lo sumo, una petición por segundo), y otra a nivel de cuota, que es el número de peticiones totales que pueden realizarse en un período determinado, en nuestro caso, un mes. Esta cuota puede aumentarse si se contrata una cuenta de usuario profesional, mediante un plan de precios. No obstante, en varias ocasiones, la empresa nos ha elevado la cuota tras solicitud por nuestra parte, sin coste alguno, debido al carácter académico, y no profesional, de este trabajo.

Para llevar a cabo estas limitaciones, la empresa utiliza un servidor *proxy*, que es encargado de imponer las restricciones. En concreto, utiliza *Apigee*, del que puede encontrarse más información en su sitio web: https://apigee.com/api-management/#/homepage

Este *proxy*, utiliza un sistema de penalización, que se activa en cuanto una de las restricciones es violada. La penalización consiste habitualmente en prohibir el acceso a la información durante un tiempo determinado, por ejemplo, en el caso de la violación de cuota, la penalización consiste en impedir el acceso a la base de datos de Idealista desde el momento de la infracción hasta la finalización del ciclo, normalmente, el mes en curso. (Ecma International, 2016)

La API se utiliza, básicamente, para realizar peticiones *HTTP* al servidor de Idealista mediante el proxy de Apigee, y a través de la dirección https://idealista-prod.apigee.net/public/2/search a la cual se le añaden parámetros opcionales, que permiten realizar la búsqueda de los inmuebles requeridos, siguiendo una serie de criterios establecidos.

Una vez realizada la petición, el servidor responde mediante un mensaje codificado en JSON, que contiene parte de los inmuebles resultantes, así como el número total de éstos, que coinciden con los criterios seleccionados, y el número total de páginas que contiene la información resultado de la búsqueda.

JSON, (*JavaScript Object Notation*) es un formato de texto ligero que se utiliza para el intercambio de datos. Realmente, es un subconjunto de la notación literal de objetos de

JavaScript, que, debido a su amplia adopción como alternativa a *XML*, es considerado, ya, un formato de lenguaje independiente y una referencia a utilizar en la transferencia de datos en los servicios web. Puede verse una especificación de este lenguaje en http://www.json.org/

La API impone una restricción de 50 inmuebles por página, de forma que se evite el envío de gran cantidad de información en cada una de las peticiones. Esto implica la necesidad de realizar una petición para cada una de las múltiples páginas que resultan de la búsqueda realizada, lo que tiene un efecto negativo en el consumo de la cuota mensual. Un ejemplo de la información que devuelve la API puede en la *Figura 52*, que una vez formateado queda como se muestra en la Figura 53.


```
{"actualPage":1,"elementList":[{"address":"zamorano","agency":false,"agentLogo":"","ba
throoms":0,"condition":"","country":"es","description":"","distance":"289","district":
"Centro - Casco
Histórico","exterior":false,"favComment":"","favourite":false,"floor":"","hasVideo":fa
lse,"latitude":37.8899812,"longitude":-
4.7751438,"municipality":"Córdoba","neighborhood":"Casco Histórico - Ollerías -
Marrubial","numPhotos":2,"operation":"V","photosUrl":"www.idealista.com/inmueble/34561
302","price":19000,"propertyCode":"34561302","propertyType":"garage","propertyTypeCode
":"G","province":"Córdoba","region":"","rooms":0,"showAddress":false,"size":0,"subregi
on":"","thumbnail":"https://img3.idealista.com/blur/WEB_LISTING/0/id.pro.es.image.mast
er/17/1f/27/166062926.jpg","url":"www.idealista.com/34561302","userCode":"","userType"
:0,"videoType":"F"},{"address":"barrio
centro","agency":false,"agentLogo":"","bathrooms":0,"condition":"","country":"es","des
cription":"","distance":"178","district":"Centro - Casco
Histórico","exterior":false,"favComment":"","favourite":false,"floor":"bj","hasVideo":
false,"latitude":37.8893015,"longitude":-
4.7765453,"municipality":"Córdoba","neighborhood":"Centro","numPhotos":2,"operation":"
V","photosUrl":"www.idealista.com/inmueble/27695533","price":22000,"propertyCode":"276
95533","propertyType":"garage","propertyTypeCode":"G","province":"Córdoba","region":""
,"rooms":0,"showAddress":false,"size":0,"subregion":"","thumbnail":"https://img3.ideal
ista.com/blur/WEB_LISTING/0/id.pro.es.image.master/99/cc/8b/92015041.jpg","url":"www.i
dealista.com/27695533","userCode":"","userType":0,"videoType":"F"}],"itemsPerPage":50,
"lowerRangePosition":0,"numPaginations":0,"paginable":false,"resultSummary":"Comprar,
Garajes, De todos los precios, De todos los
tamaños","total":2,"totalPages":1,"upperRangePosition":2}
```


Figura 52. Ejemplo de la información devuelta por la API. Sin formatear. (Fuente: Idealista.com)

```
{
  "actualPage": 1,
  "elementList": [
    {
      "address": "zamorano",
      "agency": false,
      "agentLogo": "",
      "bathrooms": 0,
      "condition": "",
      "country": "es",
      "description": "",
      "distance": "289",
      "district": "Centro - Casco Histórico",
      "exterior": false,
      "favComment": "",
      "favourite": false,
      "floor": "",
      "hasVideo": false,
      "latitude": 37.8899812,
      "longitude": -4.7751438,
      "municipality": "Córdoba",
      "neighborhood": "Casco Histórico - Ollerías - Marrubial",
      "numPhotos": 2,
      "operation": "V",
      "photosUrl": "www.idealista.com/inmueble/34561302",
      "price": 19000,
      "propertyCode": "34561302",
      "propertyType": "garage",
      "propertyTypeCode": "G",
      "province": "Córdoba",
      "region": "",
      "rooms": 0,
      "showAddress": false,
      "size": 0,
      "subregion": "",
      "thumbnail": "https://img3.idealista.com/blur/WEB_LISTING/0/id.pro.es.image.master/17/1f/27/166062926.jpg",
      "url": "www.idealista.com/34561302",
      "userCode": "",
      "userType": 0,
      "videoType": "F"
    },
    {
      "address": "barrio centro",
      "agency": false,
      "agentLogo": "",
      "bathrooms": 0,
      "condition": "",
      "country": "es",
      "description": "",
      "distance": "178",
      "district": "Centro - Casco Histórico",
      "exterior": false,
      "favComment": "",
      "favourite": false,
      "floor": "bj",
      "hasVideo": false,
      "latitude": 37.8893015,
      "longitude": -4.7765453,
      "municipality": "Córdoba",
      "neighborhood": "Centro",
      "numPhotos": 2,
      "operation": "V",
      "photosUrl": "www.idealista.com/inmueble/27695533",
      "price": 22000,
      "propertyCode": "27695533",
      "propertyType": "garage",
      "propertyTypeCode": "G",
      "province": "Córdoba",
      "region": "",
      "rooms": 0,
      "showAddress": false,
      "size": 0,
      "subregion": "",
      "thumbnail": "https://img3.idealista.com/blur/WEB_LISTING/0/id.pro.es.image.master/99/cc/8b/92015041.jpg",
      "url": "www.idealista.com/27695533",
      "userCode": "",
      "userType": 0,
      "videoType": "F"
    }
  ],
  "itemsPerPage": 50,
  "lowerRangePosition": 0,
  "numPaginations": 0,
  "paginable": false,
  "resultSummary": "Comprar, Garajes, De todos los precios, De todos los tamaños",
  "total": 2,
  "totalPages": 1,
  "upperRangePosition": 2
```


Figura 53. Ejemplo de la información devuelta por la API. Formateado. (Fuente: Idealista.com)

Como puede apreciarse en la Figura 53, para cada inmueble, se devuelve un conjunto de atributos que definen sus características, como son el precio, el número de habitaciones o el número de baños. No obstante, existe otro conjunto de atributos que no devuelve la API como resultado de la búsqueda, pero que pueden ser útiles para su procesamiento o estudio estadístico, como son el estado de conservación del inmueble, la disponibilidad de ascensor, cochera, piscina, etc. A este tipo de campos le denominamos campos *buscables*.

Como se ha indicado anteriormente, los campos *buscables*, no aparecen como resultado de la búsqueda de inmuebles, pero sí pueden ser usados para realizar filtros en

éstas, de tal forma que, el mecanismo para obtener información de estos campos, consiste en lanzar nuevas peticiones al servidor solicitando sólo aquellos inmuebles que cumplen una determinada propiedad *buscable*.

Este procedimiento incrementa de manera notable el número de peticiones a realizar en cada proyecto, pero nos permite obtener esta valiosa información. El listado de campos *buscables* se puede consultar en la Tabla 22.

Tabla 22. Tabla de campos buscables.

Campo *Buscable*	Aplicable a:	Valores posibles que pueden tomar		
COCHERA	Vivienda Oficina	Verdadero	Falso	
ASCENSOR	Vivienda Oficina	Verdadero	Falso	
PISCINA	Vivienda	Verdadero	Falso	
AIRE ACONDICIONADO	Vivienda Oficina Local	Verdadero	Falso	
TERRAZA	Vivienda	Verdadero	Falso	
TRASTERO	Vivienda	Verdadero	Falso	
TENDEDEROS	Vivienda	Verdadero	Falso	
ARMARIOS EMPOTRADOS	Vivienda	Verdadero	Falso	
ESTADO	Vivienda	Bueno	Nuevo	Reformado
LOCALIZACIÓN	Local	A pie de calle	Centro comercial	Entreplanta
ESQUINA	Local	Verdadero	Falso	
SALIDA DE HUMOS	Local	Verdadero	Falso	
AGUA CAL. INDEP.	Oficina	Verdadero	Falso	
CALEFACCIÓN INDEP.	Oficina	Verdadero	Falso	
SEGURIDAD	Oficina	Verdadero	Falso	
DISTRIBUCIÓN	Oficina	Abierta	Cerrada	
USO	Oficina	Compartido	Exclusivo	
PUERTA AUTOMATICA	Cochera	Verdadero	Falso	
PARKING MOTO	Cochera	Verdadero	Falso	

Fuente: Elaboración propia a partir de los datos recogidos por Idealista

Se ha optimizado el proceso de obtención de los campos *buscables*, eliminando del proceso de búsqueda, aquellas duplas campo-valor que no aportan información, o cuyo resultado es el mismo que el obtenido al no realizar dicho filtrado.

Por tanto, para la obtención de la información, se realiza una primera búsqueda con los criterios básicos: el centro geográfico, la distancia, el tipo de inmueble (piso, local,

cochera u oficina), y el tipo de operación (venta o alquiler), y, a continuación, una búsqueda, con sus páginas correspondientes, para cada campo-valor relevante.

6.3.1.2. La base de datos

La base de datos juega un papel muy importante en el funcionamiento del programa, ya que mitiga las limitaciones de la API relativas a la cuota, almacenando toda la información que se obtiene a través de las búsquedas solicitadas a la base de datos de Idealista.

Se ha escogido la base de datos H2, para mejorar la portabilidad de la aplicación, ya que ésta puede distribuirse con la aplicación, y no requiere de ningún tipo de instalación adicional ni de configuraciones que hagan necesario el aporte de un administrador. Además, al ser multiplataforma, no limita su uso a ningún tipo de sistema operativo. Para más información acerca de H2 consultar http://www.h2database.com/html/main.html

Esta base de datos está completamente desarrollada en Java y dispone de dos modos de funcionamiento: embebido y en modo servidor. El modo escogido es el primero de ellos, por su mayor simplicidad y eficiencia respecto al modo servidor, que, aunque permite la realización de peticiones recurrentes, debido a las restricciones impuestas por la API en cuanto al número de peticiones simultáneas, esta funcionalidad no puede ser aprovechada.

El esquema de base de datos utilizado para modelar la información obtenida es el que se muestra en la Figura 54.

Cuando aparece un nuevo inmueble en la búsqueda, éste se crea en la base de datos. En las sucesivas actualizaciones que se hacen del proyecto, se actualizan los datos del inmueble sobrescribiendo el valor inicial, salvo el relativo al precio, cuyo nuevo valor se incluye en un nuevo registro.

A continuación, se detallan las tablas del diagrama entidad-relación mostradas en la Figura 54.

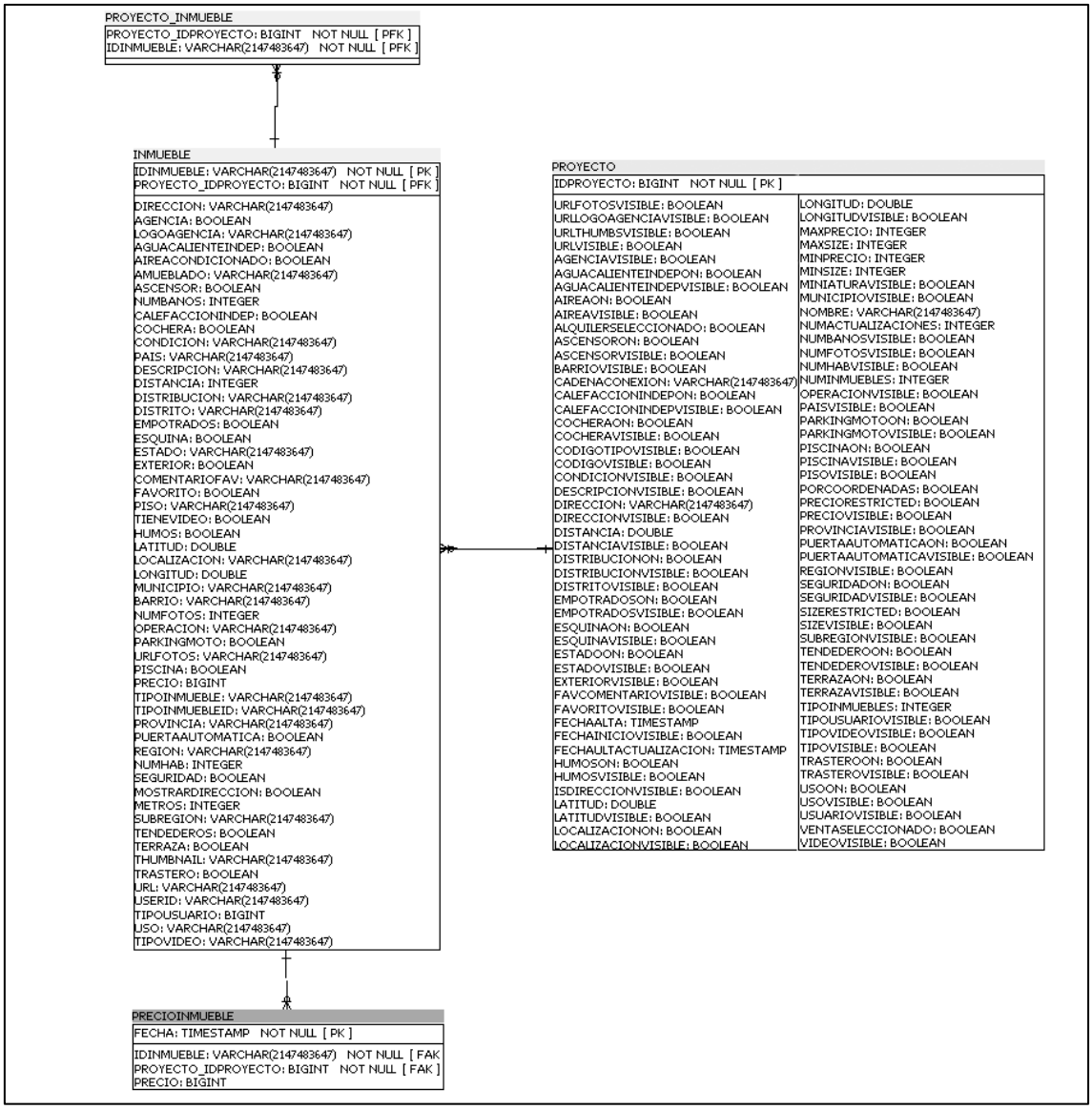


Figura 54. Esquema de la base de datos.

- Tabla *Proyectos*. Almacena la información relativa a los proyectos que se crean en la aplicación, y contiene todos los detalles del mismo. Almacena las coordenadas geográficas, el radio de búsqueda utilizado, el tipo de inmueble y de operación que se realiza con ellos (venta o alquiler) y si tuviera algún filtrado en cuanto al precio o al tamaño (mínimo o máximo).

  En cada proyecto, puede seleccionarse, además, el número de campos *buscables* que quieren utilizarse en la búsqueda durante la vida útil del proyecto, así como la configuración de los campos que quieren mostrarse en los resultados y en las exportaciones de datos. Estas configuraciones, también son almacenados en la tabla de proyectos.

- Tabla *Inmueble*. Almacena toda la información referente a los inmuebles pertenecientes al proyecto. Esta tabla contiene todos los campos recuperados en la búsqueda, incluidos los *buscables*, y el precio base del inmueble, que es el primer precio que se guarda de él, es decir, el que resultó de la primera búsqueda en la que se encontró dicho inmueble.

- Tabla *PrecioInmueble*. Esta tabla almacena la evolución de precios que ha tenido cada inmueble desde la primera actualización hasta la última realizada. Hay que tener en cuenta, que los inmuebles pueden tener diferente número de registros de precios, ya que, si éste desaparece de la base de datos de Idealista, no volverá a almacenarse información relativa a su precio. Además, si un inmueble aparece por primera vez en alguna actualización posterior a la inicial, sólo se dispondrá de información de su precio desde ese instante.
- Tabla *Proyecto_Inmueble*. Es la tabla que relaciona los inmuebles con los proyectos, ya que indica qué inmueble pertenece a cada uno de los proyectos existentes. Es la abstracción lógica de la creación de la relación "*muchos a muchos"* existente entre la tabla Proyectos y la tabla Inmuebles.

6.3.1.3. Lógica de Adquisición

Existen dos procesos que realizan la adquisición de datos. El primero de ellos es el que crea el proyecto con las características de filtrado que se desean configurar, y el segundo es el proceso que se utiliza para la actualización de la información de los inmuebles del proyecto.

Para optimizar el número de peticiones, en el proceso de creación del proyecto sólo se muestra una *previsualización* del resultado de la búsqueda, que contiene los primeros 50 inmuebles. En este momento no se almacena información de ningún inmueble, de ello se encarga el proceso de actualización descrito a continuación.

El funcionamiento del proceso de actualización de la información relativa a los inmuebles es el siguiente:

1. Se lanza una primera búsqueda con los parámetros de filtrado básicos: centro geográfico, distancia, precio y tamaño; limitando la búsqueda a 1 inmueble de forma que obtengamos el número total de inmuebles de la búsqueda ordenado descendentemente por el precio.
2. Tras numerosas pruebas, identificamos una nueva limitación de la búsqueda: Si el número de páginas de la búsqueda es superior a 50, es decir, si el número de inmuebles supera los 2500 (50 páginas con 50 inmuebles por página como máximo), la información devuelta a partir de la página 51 es errónea, ya que en realidad se obtiene, nuevamente, información de una de las 50 primeras páginas elegidas por el sistema al azar. Por este motivo, se debe subdividir el proceso de adquisición de datos, de forma que los resultados puedan ser mostrados en un número de páginas igual o menor de 50. Para realizar esta subdivisión de las búsquedas, se utiliza un procedimiento de bipartición en función del precio. En caso contrario, el proceso continuaría en el punto 3.

2.1. Se *deserializa* el inmueble obtenido en la búsqueda del punto 1, que consiste en convertir la salida, en formato JSON, en la representación en Java del objeto tipo Inmueble, para la utilización de sus datos y para su posterior almacenamiento en la BBDD. Para facilitar el uso de JSON, se ha utilizado la librería *google-gson* en su versión 2.3.1, debido a que es una librería ampliamente conocida y de código abierto. (Singh, Leitch y Wilson, 2016) Puede obtenerse más información de la librería en su sitio web de Github: https://github.com/google/gson

2.2. Se calcula la media entre el límite mínimo de precio y el de menor valor entre: el límite máximo establecido en el proyecto y el precio devuelto en la búsqueda obtenida en el paso anterior, que es el de mayor valor de la búsqueda por estar ordenado de forma descendente. Se almacena la información del punto medio.

2.3. Se divide la búsqueda en dos búsquedas que contienen un menor número de inmuebles, al restringir los precios mínimo y máximo de la siguiente forma: un primer tramo de precios que va desde el límite mínimo de precio hasta el punto medio, y un segundo, desde el punto medio hasta el límite máximo de precio. Para estas dos búsquedas, se repite el proceso, volviendo a dividir en tramos de precios más pequeños, en caso de que sea necesario. Este proceso se repite hasta que el número de inmuebles en cada tramo de precios sea inferior a la limitación impuesta, 2500 inmuebles.

2.4. Una vez obtenidos los tramos en los que se divide la búsqueda, se optimiza su uso para reducir el número de peticiones. Para ello, si al sumar el número de inmuebles de dos tramos consecutivos, el resultado es un número inferior a 2500, se unen los dos tramos en uno sólo, y, por tanto, las dos búsquedas en una. Esto es importante, sobre todo en los tramos que contienen precios muy altos, donde suele ser habitual encontrar un menor número de inmuebles en cada tramo.

3. A continuación, se realiza la primera búsqueda de información, sin establecer ningún tipo de restricción dada por los campos buscables, de forma que aparezcan todos los inmuebles contenidos en ella. Esto se realiza en todos los tramos, si éstos han sido creados, y todas y cada una de las páginas resultantes de cada tramo, deserializando los inmuebles y guardándolos en una lista para su posterior almacenamiento en la BBDD.

4. A continuación, se itera a través de todos los criterios buscables de que disponga el proyecto, según su tipo (locales, viviendas, garajes u oficina) y según los criterios seleccionados por el usuario en la creación del proyecto.

4.1. Para cada campo buscable, se itera, a su vez, en cada uno de los valores posibles que éste pueda tomar, los cuales pueden consultarse en la Tabla 23, teniendo en cuenta que se ha aplicado la optimización antes mencionada de forma que se eviten búsquedas innecesarias. Si en la búsqueda principal no se hubieran generado tramos de precios, se omitirá la búsqueda de tramos, en este punto, para evitar peticiones innecesarias.

4.2. Los inmuebles que aparecen en las búsquedas, son aquellos que cumplen la propiedad campo-valor seleccionada en cada iteración, cuyo número será igual o inferior, normalmente inferior, al número obtenido en el paso 3. En este punto se lleva a cabo la obtención de tramos de la misma manera que se ha realizado en el punto 2, aplicado al criterio concreto.

4.3. Se deserializan los inmuebles resultantes y se actualiza la lista obtenida en el paso 3, para añadir la información resultante del nuevo criterio.

4.4. Se repite el proceso para cada uno de los campos-valores.

5. Por último, y una vez que el procedimiento se ha completado de manera correcta con todos los campos buscables, los datos resultantes se almacenan en la base de datos.

Para mejorar el proceso de almacenamiento de la información, modificación y borrado de la información, se ha hecho uso de la librería *Eclipselink* en su versión 2.5.0.

Eclipselink es un framework extensible que permite a un desarrollador interactuar con varios servicios, como bases de datos. Soporta varios estándares de sistemas de persistencia, incluido Java Persistence API (JPA) que es el que se ha utilizado en este proyecto. Esta librería es la encargada de realizar el proceso de persistencia, es decir, de almacenamiento y gestión, de la información desde y hacia la base de datos. (The Eclipse Foundation, 2015)

Hay que tener en cuenta si el proceso es una actualización posterior de un proyecto ya existente, y no su primera actualización, ya que, si un inmueble no se encontraba previamente en la base de datos, éste se añade según el proceso descrito. Pero si ya estaba incluido en ésta, se actualiza su información por si alguna característica del mismo ha cambiado, y se crea una nueva entrada de precio para el inmueble, sin sobrescribir la anterior, e incluyendo la fecha en la que este precio ha sido obtenido. De esta manera, se registra la evolución del precio de los inmuebles a lo largo del tiempo.

6.3.1. Interfaz de usuario

Es el sistema encargado de la comunicación entre el usuario y la aplicación. Es el nexo de unión entre los otros dos sistemas.

Permite la visualización de los datos recogidos en el proceso de adquisición de datos, así como el filtrado de los mismos o su exportación a ficheros CSV para su uso en otros programas de cálculo o estadísticos.

Para el sistema de filtrado se ha hecho uso de unas de las librerías graficas más reconocidas en *JavaFX*: *Controlsfx*, en su versión 8.40.12 (Giles, 2016). Tiene un nivel tal de implantación, que parte de su código ha sido recogido ya, como parte Java, a partir de la versión 1.8.0_40, motivo por el cual, su instalación, o una versión superior de ésta,

es un requisito indispensable para el correcto funcionamiento del software. Puede encontrarse más información de dicha librería en http://fxexperience.com/controlsfx/

El manejo de los ficheros CSV y la exportación de los datos se lleva a cabo haciendo uso de la librería *opencsv*, concretamente la versión 3.5. La elección de esta librería está basada en la naturaleza de código abierto de la misma. (Slashdot Media, 2017) Se puede encontrar la página del proyecto en http://opencsv.sourceforge.net/

Además, se han implementado todos los cuadros de diálogo necesarios para la creación de los proyectos, la visualización de los datos y la realización de los estudios estadísticos que pueden ser llevados a cabo, así como la exposición de sus resultados.

La salida obtenida en cada uno de los estudios estadísticos, se muestra en formato HTML, por ser éste un lenguaje portable y, por tanto, poder visualizarse en cualquier navegador que soporte JavaScript, la inmensa mayoría de ellos. No obstante, y para asegurar su compatibilidad, se recomienda el uso de navegadores que soporten la versión 5 de HTML.

Para facilitar la creación de los ficheros HTML resultantes de los estudios estadísticos se ha optado por usar *jsoup*, librería de código abierto, que implementa un *parser*, o analizador sintáctico, y que, a través de sus funciones, permite recrear código HTML de una manera más limpia y eficiente que si se utilizara funciones de composición de cadenas. (Hedley, 2017) Se puede encontrar información sobre éste en https://jsoup.org/

6.3.2. Sistema de análisis de información

El sistema de análisis de información es el encargado de tratar la información almacenada a través del sistema de adquisición de datos. Este sistema está basado principalmente en tres capas: la lógica de programa, que comunica con la interfaz de usuario, *JRI*, y el motor de R.

JRI es una interfaz Java/R que permite ejecutar aplicaciones R dentro de aplicaciones Java en un único hilo. Básicamente, carga la librería dinámica de R dentro de Java y provee a éste de funcionalidades, que permiten el lanzamiento de sentencias en lenguaje R contra el motor del mismo.

La interfaz JRI forma parte de la librería *rJava*, que está destinada a la integración de Java con R, permitiendo ejecutar código Java dentro del propio motor de R. Del mismo modo, a través de JRI se consigue la integración inversa.

Es necesario destacar que JRI usa código nativo, lo que significa que parte del mismo está desarrollado con otro lenguaje que no es Java y que por tanto no es multiplataforma, por lo que requiere que en el sistema operativo donde vaya a ejecutarse la aplicación, sea cargada una librería dinámica que sea compatible con dicho sistema.

Esto podría limitar las características multiplataforma de la aplicación. Afortunadamente, JRI está disponible en multitud de sistemas operativos, como Windows, Mac OS X, Sun y Linux, en versiones tanto de 32 como de 64 bits.

Para la utilización de JRI deben verificarse dos requisitos: disponer de las librerías nativas de R y de JRI, y tener la capacidad de enlazarlas. La librería de R se encuentra en la carpeta de instalación del propio programa, mientras que la librería de JRI puede obtenerse a través de la instalación del paquete rJava desde los repositorios de R.

Por tanto, previo al proceso de instalación de la aplicación, el usuario deberá disponer de cualquiera de las versiones de R soportas por la aplicación, comentadas anteriormente, así como de la librería rJava instalada en el mismo.

Durante el proceso de instalación se realiza la configuración necesaria para la verificación del segundo requisito, consistente en enlazar las librerías nativas a la aplicación. La especificación de ésta debe realizarse de forma manual, como se explicará posteriormente.

La versión de rJava utilizada es la 0.9-8, que es la disponible en el momento de la redacción de este trabajo, a través de los repositorios de R Cran, (Urbanek, 2016), o través de la página oficial del proyecto https://rforge.net/rJava/

La versión de JRI que lleva integrada rJava es la 0.5-0 de la que puede obtenerse más información en https://rforge.net/JRI/

JRI dispone de una serie de clases y métodos (API) que facilitan la utilización de las sentencias y funciones de R. Por ello, también debe disponer de un mecanismo de conversión de datos que permita el envío y la recogida de información de R.

Para cada uno de los estudios que se deseen realizar, el usuario deberá seleccionar, a través de la interfaz, los datos y las variables con los que trabajar, así como los parámetros del estudio.

Antes de realizar la petición, el sistema debe preparar la información contenida en la base de datos y transformarla, para su procesamiento por el motor de R. Han sido definidas en JRI, un conjunto de clases que son compatibles con los tipos de datos de R y que pueden ser enviadas a través del mismo. Se realiza, entonces, la conversión de datos entre Java y R a través de dichas clases, generando un conjunto de variables.

Una vez hecho esto, se envían dichas variables al motor de R, de modo que se pueda realizar la llamada a la función correspondiente al estudio que se está realizando, y comprobando, previamente, los posibles errores que pudieran generarse en la introducción de los parámetros por parte del usuario.

En ese momento, el motor de R toma el control y trabaja con la función seleccionada. Devuelve como resultado de la ejecución, un vector que contendrá los resultados de los estudios, los posibles errores que se hayan podido producir, y las rutas de los gráficos que han sido generados.

Una vez recibido el vector de resultados de R, se vuelve a realizar la conversión de los datos, en este caso en sentido inverso, para su tratamiento en Java y su posterior presentación a través de la interfaz de usuario.

Una vez realizada la conversión, se iterará a través del vector de valores obtenidos, para transformarlos a su salida en HTML, de forma que pueda ser mostrado en el panel de resultados, lo que permite su visualización.

6.4. Tipos de inmuebles en estudio y características disponibles

Como se ha comentado anteriormente, InmoDataAnalizador permite la creación de diferentes tipos de proyectos según la tipología del inmueble que interese estudiar y la operación a realizada con el mismo. Existe un conjunto de variables resultado del proceso de búsqueda que son comunes a todos los inmuebles, pero otras variables son específicas de cada proyecto.

En primer lugar, se detallan los tipos de inmuebles disponibles para el estudio, y sus operaciones, y, por último, las variables que pueden ser analizadas en cada uno de ellos.

6.4.1. Tipos de inmuebles y operaciones

El sistema de búsqueda y almacenamiento de inmuebles disponible en el software propuesto en este trabajo, permite realizar el estudio de cuatro tipos de inmuebles:

- Viviendas: Incluye diferentes tipos de inmuebles destinados para que habiten personas, tanto las que se encuentran en un edificio con un mayor número de ellas, como las unifamiliares, que son aquellas en las que una sola familia ocupa el edificio al completo.

- Locales: Son los inmuebles destinados a realizar una actividad económica. Estos incluyen los locales comerciales y las naves industriales, con características de tamaño, distribución y servicios diferentes. Los locales suelen ser pequeños inmuebles situados en la planta baja o la entreplanta de un edificio, que tienen acceso directo a la calle que normalmente se encuentran en zonas comerciales ya que se utilizan para actividades en las que el contacto con el cliente es continuo. Las naves industriales, por el contrario, suelen ser inmuebles de gran tamaño situados en zonas industriales con buenos accesos por carretera, tienen una zona, la de mayor tamaño, normalmente diáfana, en la que se realiza la actividad o se almacena material, y una más pequeña con una o varias oficinas donde se realiza el trabajo administrativo.

- Oficinas: Son los espacios destinados a la realización de actividades profesionales que no necesariamente requieren de un contacto permanente con el cliente, por lo que no necesitan estar situadas a pie de calle. Están situados en edificios que incluye en exclusiva oficinas, o en otros en los que también hay viviendas. Su situación geográfica en la ciudad suele estar muy acotada en zonas específicas, como centros financieros.

- Cocheras: Son espacios reservados para el estacionamiento de vehículos y, a veces, incluye un pequeño almacén. Puede ser de uso individual y, por tanto, un inmueble independiente, o, por el contrario, ocupar un espacio dentro de un garaje comunitario. En ambos escasos, su espacio es reducido.

Se contemplan, a su vez, dos tipos de operaciones:

- Venta. Es la acción de traspasar la propiedad de un bien tras el pago de una cantidad de dinero acordada.
- Alquiler. Es la acción de pagar una determinada cantidad de dinero para hacer uso de un inmueble, durante un período de tiempo determinado. Este acto suele estar definido por el contrato de alquiler.

Por lo tanto, pueden realizarse 8 combinaciones de estudios diferentes:

- Venta de viviendas
- Alquiler de viviendas
- Venta de locales
- Alquiler de locales
- Venta de oficinas
- Alquiler de oficinas
- Venta de cocheras
- Alquiler de cocheras

6.4.2. Variables de estudio

La búsqueda incluye la información de un conjunto de características que dependen del tipo de inmueble buscado, pero no varía al cambiar el tipo de operación realizada. Alguna de las variables recogidas, no ofrecen información alguna, debido a que el acceso a los datos puesto a nuestra disposición de forma generosa por la empresa Idealista S.A., es un acceso restringido a un servicio de pago que esta empresa ofrece a otras empresas del sector, como inmobiliarias.

6.4.2.1. Variables comunes a todos los proyectos

Los campos comunes a todos los tipos de inmuebles son los siguientes:

- *Fecha Alta*: Es la fecha en la que el inmueble ha sido incorporado a la base de datos del programa. -No es la fecha de incorporación al portal Idealista -. Tiene el formato de fecha: dd/mm/aaaa, es decir, el día, seguido del mes y del año en que se incluyó el inmueble en la base de datos del programa.
- *Longitud*: Es el ángulo entre el meridiano de Greenwich y el meridiano que pasa por la posición geográfica del inmueble. Se expresa en grados, en forma incompleja e indicando la situación respecto al meridiano de referencia mediante signo, de forma que, si éste es positivo, el inmueble se encuentra al este del meridiano de Greenwich, y si es negativo, al oeste del mismo. Los valores posibles para el territorio nacional son los comprendidos entre -18.160556 (situado en el municipio de La Frontera, en la isla de Hierro) y 4.329444 (situado en la isla Menorca).
- *Latitud*: Es el ángulo entre el plano ecuatorial y el paralelo que pasa por la posición geográfica del inmueble. Se expresa en grados, en forma incompleja e indicando la situación respecto al eje de referencia, mediante signo. Si su valor es positivo, el inmueble se encuentra en el hemisferio norte de la Tierra, y si es negativo, en el sur. Todos los puntos geográficos de España, y, por tanto, todos los inmuebles tienen latitud positiva comprendida entre 27.637778 (al sur de la isla de El Hierro), y 43.793889 (en el extremo norte de la provincia de La Coruña).
- *Distancia*. Esta variable mide, en metros, la distancia, en línea recta, entre el centro de la búsqueda y la posición del inmueble en cuestión.
- *Con Dirección*. Esta variable es de tipo lógico, de forma que el valor *true* significa que el inmueble dispone en la ficha de la información relativa a la dirección geográfica exacta. El valor *false* indica que no dispone de esta información. No obstante, esta variable no es considerada en el estudio por devolver siempre el valor false, aun disponiendo el inmueble de dirección completa.
- *Dirección*. Se refiere a la dirección exacta en la que se encuentra situado el inmueble en la ciudad. Hay ocasiones en las que la información es completa, incluyendo número de calle, otras en las que identifica la calle en la que se encuentra, pero no el número de la calle en el que se encuentra, y las que sólo indican la zona o barrio en la que están situadas, por expreso deseo de la persona que ha publicado el anuncio.
- *Barrio*. Indica el vecindario o barrio en el que se encuentra situado el inmueble. El barrio indicado por esta variable no necesariamente coincide exactamente con los barrios que administrativamente son considerados por los ayuntamientos de los municipios. Unas veces hace referencia a zonas geográficas que engloban varios barrios o varias partes de éstos, y en otras, la zona se delimita con mayor exactitud a entornos concretos.
- *Distrito*. Indica una zona geográfica, más amplia de la ciudad que el barrio, en la que se encuentra situado el inmueble. Los distritos empleados por el portal inmobiliario no necesariamente coinciden con los determinados por el ayuntamiento.
- *Municipio*. Es la circunscripción administrativa en la que se encuentra situado el inmueble. Los resultados posibles dependerán del ámbito geográfico en el que se realice la búsqueda.

- *Provincia*. Provincia en la que se encuentra situada el inmueble. La búsqueda puede ser realizada. Como la búsqueda puede realizarse en cualquier lugar del territorio nacional, el valor de esta variable puede coincidir con cualquiera de las existentes.
- *Región*. Variable diseñada para indicar otra demarcación superior a la provincia. No disponible en ninguna de las búsquedas realizadas Indica la comarca o región en la que se encuentra el inmueble.
- *Subregión*. Parte de la región en la que se encuentra situado el inmueble, en caso de estar dividido en subregiones. Tampoco se ha obtenido valor alguno en ninguna de las búsquedas realizadas.
- *País*. Es un código de dos letras que indica el país al que pertenece el inmueble. Por lo que siempre toma el valor *es*. Esta variable justifica su existencia en que el portal inmobiliario idealista.com dispone de oferta de inmuebles en Portugal e Italia, y es intención de la empresa, ampliar la consulta a estos países, para los que los valores de la variable serán, respectivamente, *pt* e *it*.
- *Código*: Es un valor numérico, de entre 7 y 8 dígitos, que identifica, de forma única, cada inmueble de la base de datos y que será utilizado por el programa para identificar los inmuebles en los estudios estadísticos en los que esto sea necesario, como en la identificación de observaciones influyentes.
- *Operación*: Esta variable indica con una única letra si la operación es de venta (V) o de alquiler (A). Como en la configuración de la búsqueda se discrimina entre una de estas dos opciones, esta variable tomará el mismo valor para todos los inmuebles de un proyecto.
- *Tipo*: Esta variable indica el tipo de inmueble en estudio. En el caso de locales, oficinas y cocheras, tomará un mismo valor para todos los inmuebles: *premise*, *office* y *garage* respectivamente. En el caso de viviendas, esta variable puede tomar seis valores distintos:
    - El valor *chalet*, para viviendas unifamiliares que está situado en una finca que dispone también de terreno sin construir como puede ser un jardín o un patio.
    - El valor *countryHouse* sirve para identificar casas de campo o rurales, que son viviendas unifamiliares con edificio propio que se enmarcan en un entorno rural.
    - El valor *duplex* hace referencia a viviendas unifamiliares compuestas por dos plantas unidas por una escalera interior. A diferencia de los chalés, los dúplex no necesariamente disponen de un terreno sin edificar en la finca.
    - El valor *flat* hace referencia a pisos o apartamentos, que son viviendas distribuidas en una sola planta, de menor superficie que los chalés o las casas de campo, y que situadas en el interior de edificios comunes que están compuestos por varias viviendas.
    - El valor *penthouse* sirve para identificar las viviendas situadas en la planta superior de un edificio de viviendas. Los áticos suelen ser viviendas con unas calidades superiores a las del resto del edificio, que disponen de grandes vistas y, además, de una zona al aire libre de recreo en la azotea.

    - El valor *studio* se utiliza para identificar las viviendas que disponen de una sala multifuncional que sirve de comedor, sala de estudio o trabajo y dormitorio; una cocina y un baño, que es el único habitáculo separado de la sala principal.

- *Cod. Tipo*. Esta variable, código tipo, identifica el tipo de vivienda indicado en la variable anterior, mediante un código formado por una o dos letras, que son:
    - L para locales.
    - O para oficinas.
    - G para cocheras o garajes.
    - Para las viviendas los códigos están formados por dos letras, la primera de ellas es V que indica que es una vivienda, la segunda letra indica el tipo. Por tanto: VI: chalé o vivienda independiente, VD: dúplex, VP: piso, VA: ático, VE: estudio. Cuando la vivienda es una casa rural, este campo se queda vacío.
- *Descripción.* Esta variable es de tipo alfanumérica e incluye la descripción del inmueble dada por el ofertante. No obstante, en la base de datos aparece este campo vacío en todos los casos, ya que la empresa no provee este dato.
- *Num Hab*. La variable que indica el número de habitaciones de las que se compone el inmueble es de tipo numérico y está disponible para todos los tipos de viviendas, aunque sólo mide esta característica en las viviendas. Para locales, oficinas y cocheras toma el valor 0 en todos los casos.
- *Num Baños.* Es una variable numérica que indica el número de baños y aseos que tiene el inmueble, salvo que éste sea una cochera, para la que se obtiene siempre el valor 0.
- *Tamaño.* Es, también, de tipo numérico, y representa la superficie, en metros cuadrados, del inmueble. Para el tipo cochera, esta variable, siempre toma el valor 0.
- *Piso*. Indica la posición en altura o planta del inmueble dentro del edificio. Puede tomar valores numéricos para identificar el número de planta en la que se encuentra, y códigos no numéricos como bj para indicar que se encuentra en el bajo del edificio, en, para indicar que se encuentra en la entreplanta (planta situada entre el bajo y la primera planta), ss, para el semisótano y st para los inmuebles situados en el sótano.
- *Precio.* Esta variable indica el valor en euros solicitado por el ofertante para la compra o alquiler del inmueble. Es, por tanto, el valor numérico que otorga el vendedor al inmueble, una vez valoradas las características que lo definen y la situación actual del mercado. Este campo es obligatorio para la publicación de un inmueble, por lo que todos disponen de éste en la base de datos.
- *Exterior*. Es de tipo lógico, e indica si el inmueble en estudio es exterior, es decir, si las habitaciones principales tienen ventanas a la vía pública y no a un patio interior, por lo que gozan de una mayor luminosidad natural (true), o si por el contrario es interior (false). No obstante, esta variable toma el valor false en todos los inmuebles para los que se han realizado búsquedas, por lo que no será considerada en los estudios.

- *Num Fotos.* Esta variable indica el número de fotografías aportadas por el anunciante en el anuncio del inmueble, en el portal idealista.es. Estas fotografías son las que cualquier usuario del portal podrá ver del inmueble, al acceder a su ficha.
- *URL Fotos.* Contiene la dirección web del inmueble dentro del portal inmobiliario. Es de la forma www.idealista.com/inmueble/código, donde el código es el valor numérico que identifica de forma única a cada inmueble. Al hacer doble clic en la vista de datos del programa, éste abrirá una ventana del navegador y accederá al url indicado por esta variable. Si el inmueble no está ya disponible en el portal, la web a la que dirige indicará esta situación. Inicialmente, el objetivo de la empresa propietaria del portal era incluir, en esta variable, únicamente la galería de fotografías del anuncio.
- *Url.* Contiene, al igual que en la variable anterior, un enlace al portal inmobiliario. En este caso, la dirección a la que accede es www.idealista.com/código. Durante un tiempo, al intentar acceder a esta dirección, automáticamente era redirigido a la dirección www.idealista.com/inmueble/código. No obstante, en las últimas pruebas realizadas, el portal arrojaba un mensaje indicando que la dirección no era válida.
- *Miniatura.* Es, también, una variable que contiene un enlace web para cada inmueble, pero en este caso el destino de este enlace es la página que incluye las miniaturas de las fotografías contenidas en el anuncio.
- *Agencia.* Esta variable, de tipo lógico, indica si el inmueble es ofertado por una agencia de la propiedad inmobiliaria (true) o si, por el contrario, es ofertado por un particular (false). En todas las búsquedas realizadas, el resultado ha sido el mismo: false. Sin embargo, se ha comprobado que, en numerosas ocasiones, la oferta pertenecía a una inmobiliaria, por lo que esta información no está disponible y no será utilizada.
- *Logo Agencia.* En caso de que la variable anterior identificase algún anuncio perteneciente a una agencia inmobiliaria, en este campo se incluiría la dirección web que contiene el logotipo de ésta. No obstante, al no hacerlo, este campo siempre aparece vacío.
- *Vídeo.* Esta variable indica si el anuncio dispone de vídeo, además de las habituales fotografías. Es una variable lógica que toma el valor true si dispone de vídeo y false si no dispone de éste. Tampoco está disponible esta información ya que siempre toma el valor false, aunque el anuncio disponga de vídeo.
- *Tipo Vídeo.* Esta variable está enlazada con la anterior de modo que al tomar la anterior el valor false en todos los casos, ésta toma siempre el valor F, asociado a no disponer de vídeo. Los valores I y P se reservan para los vídeos realizados por Idealista y para los vídeos propios, respectivamente.
- *Usuario.* Es un código alfanumérico, con el que el portal identifica los usuarios que publican anuncios en su web. Es únicamente de uso interno de la empresa, por lo que siempre aparece vacío.
- *Tipo Usuario.* Al igual que el anterior, es de uso interno de la empresa, por lo que no se dispone de esta información. Identifica el tipo de usuario que ha publicado cada anuncio, y siempre recibe el valor 0.

- *Comentario Fav*. Esta variable sirve para que cada usuario registrado, con acceso completo a la base de datos, añada un comentario al anuncio. En nuestras búsquedas siempre aparece vacío, ya que sólo disponemos de un acceso autorizado a la base de datos.
- *Favorito* Por el motivo explicado anteriormente, esta variable siempre toma el valor false.
- *Preciodd/mm/aa*. Esta variable, cuya denominación está formada por la palabra Precio, seguida de una fecha, indica el precio del inmueble en la actualización realizada en la fecha indicada. Al realizar la primera búsqueda se crean dos variables iguales, *Precio* y *Preciodd/mm/aaaa* con fecha la inicial. A medida que se realizan actualizaciones, se crean nuevas variables de *Precio*. Si este valor es vacío, significa que el inmueble no está aún en la base de datos o ya ha dejado de estar. También pueden aparecer distintos precios, si estos han cambiado en el tiempo.

A continuación, se muestra una tabla-resumen de los campos comunes a todos los tipos de inmuebles clasificados según definan una característica relativa al inmueble, a su localización o al anuncio disponible en el portal. (Tabla *23*)

Tabla 23. *Tabla-resumen de los campos comunes a todos los proyectos.*

Localización del inmueble	**Características del inmueble**	**Características del anuncio**
Longitud	Tipo	Fecha Alta
Latitud	Cod. Tipo	Código
Distancia	Descripción	Operación
Con dirección	Num Hab	Num fotos
Dirección	Num Baños	URL fotos
Barrio	Tamaño	Url
Distrito	Piso	Miniatura
Municipio	Precio	Agencia
Provincia	Exterior	Logo Agencia
Región	Preciodd/mm/aa	Vídeo
Subregión		Tipo Vídeo
País		Usuario
		Tipo usuario
		Comentario Fav
		Favorito

Fuente: Elaboración propia a partir de los datos recogidos por Idealista

6.4.2.2. Variables específicas de cada tipo de inmueble

Los campos aquí enumerados son los que devuelve cualquier consulta a la base de datos, independientemente del tipo de inmueble en estudio. Además de éstos, pueden realizarse consultas adicionales para obtener información añadida del inmueble, y que depende del tipo de éste que se está estudiando. Veamos a continuación las variables propias de cada uno de los tipos. Algunas de ellas son compartidas por dos o más tipos de inmuebles.

Comenzaremos por analizar las variables adicionales que pueden ser consultadas en el caso de las viviendas:

- *Cochera*. Variable de tipo lógico que informa de si la vivienda dispone de cochera individual o plaza de garaje (*true*), o por el contrario no dispone de ella (*false*).
- *Ascensor*. Nos da información de si la vivienda dispone de ascensor (*true*) o no (*false*). Esta información puede ser relevante, sobre todo si el piso en el que se encuentra la vivienda es elevado.
- *Piscina*. Informa sobre la disponibilidad (*true*) o no (*false*) de piscina individual o comunitaria con la vivienda.
- *AireAcon*. Esta variable nos indica si la vivienda está acondicionada con una máquina de aire, al menos en una de las habitaciones (*true*), o si por el contrario no dispone de ella (*false*).
- *Terraza*. Es, nuevamente, una variable de tipo lógico, que indica si la vivienda dispone de terraza (*true*) o no (*false*).
- *Trastero*. Esta variable indica si la vivienda en estudio dispone de trastero (*true*), que es una estancia que puede encontrarse en lugares anexos a la casa como son el garaje o la azotea del edificio, en el que se guardan objetos a los que se les da poco uso. Si no dispone de trastero, esta variable tomará el valor *false*.
- *Tendederos*. Esta variable indica si la vivienda dispone de una zona específica para tender la ropa. No obstante, el valor recibido por esta variable es siempre true, por lo que no se utiliza en el estudio.
- *Empotrados*. Un armario empotrado es un armario, hecho a medida y construido en el hueco de una pared, normalmente, aprovechando una zona de la habitación con poca utilidad. Esta variable nos informa si la vivienda en estudio dispone de armarios de este tipo (*true*) o no (*false*), por lo que es también una variable dicotómica.
- *Condición*. Este campo siempre es recibido vacío, pero de obtener información, ésta haría referencia al estado de amueblado del inmueble, de forma que los valores posibles son *yes* si la vivienda se encuentra amueblada, *no* si no lo está, y *Only kitchen*, si sólo dispone de cocina amueblada. Al no recibir esta información, esta variable no será utilizada, tampoco, en los estudios.

- *Estado*. La última información que es recogida acerca de una vivienda es la que hace referencia al estado de conservación de la misma. Esta variable puede tomar tres valores posibles: *good* para indicar que se encuentra en buen estado, *renew*, para indicar que necesita una reforma, y *newdevelopment* para hacer valer que es de nueva construcción.

En el estudio de locales, las variables adicionales a consultar son:

- *Localización*. Informa sobre la situación del local comercial respecto a la calle, por tanto, tomará el valor *street* si éste tiene acceso directo desde la calle y se encuentra a la misma altura que ésta, *mezzanine* si el local se encuentra en una entreplanta, y *shoppingcenter* si está en el interior de un centro comercial o edificio destinado exclusivamente a albergar locales comerciales y oficinas.
- *Esquina*. Es una variable lógica que indica si la fachada del local se encuentra en la esquina entre dos calles (*true*), lo que aumenta su visibilidad, o no (*false*).
- *Salida de Humos*. Para algunas actividades como la hostelería es de vital importancia que el local disponga de una salida de humos adecuada. Esta variable indica la existencia de esta característica (*true*) o su ausencia (*false*).
- *AireAcon*. Esta variable es compartida con las viviendas e indica si el local dispone de aire acondicionado (*true*).

Para oficinas, todas las variables opcionales son dicotómicas. Tres de ellas son compartidas con viviendas o locales. A continuación, se listan, explicando el significado de cada una de ellas:

- *Agua Caliente Indep*. La disponibilidad en la oficina de agua caliente independiente del resto del edificio es la información proporcionada por esta variable, ya que en ocasiones las calderas son comunes y en otras ni siquiera hay disponibilidad de agua caliente en la oficina. Si dispone de agua caliente de forma independiente, esta variable tomará el valor *true*, en caso contrario *false*.
- *Calefacción Indep*. Informa, al igual que la variable anterior, de la disponibilidad de forma independiente, en este caso, de calefacción.
- *Seguridad*. Esta variable indica si el edificio en el que se encuentra situada la oficina dispone de seguridad propia (*true*) o no (*false*). Esto es común en los edificios exclusivos de oficinas.
- *Distribución*. La oficina puede tener una distribución tabicada en estancias (*close*), o, por el contrario, puede ser diáfana (*open*). Esta variable nos informa de ello.
- *Uso*. La información proporcionada por esta variable hace referencia al edificio en el que se encuentra la oficina. Si es un edificio que comparte la presencia de oficinas con viviendas, esta variable toma el valor *mixed*. Si, por el contrario, sólo alberga oficinas, tomará el valor *exclusive*.
- Cochera. Al igual que en viviendas, hace referencia a la existencia de cochera asociada a la compra o al alquiler de la oficina.

- Ascensor. Informa sobre la disponibilidad de ascensor para el acceso a la planta en la que está situada la oficina.
- AireAcon. Tal y como se analizó en viviendas y locales, esta variable informa de la existencia de acondicionamiento térmico en la oficina.

Y, por último, las variables adicionales, disponibles para el tipo cochera, son únicamente tres:

- *Puerta Autom*. La apertura de la puerta del recinto en el que se encuentra la zona de estacionamiento del vehículo puede ser manual (*false*) o automática (*true*), de modo que ésta puede abrirse a distancia mediante un mando, desde el coche.
- *Parking Motoc*. Esta variable informa de si la zona de estacionamiento es de reducido tamaño y, por tanto, sólo destinada al estacionamiento de motocicletas (*true*) o si, por el contrario, sus dimensiones permiten su uso por parte de vehículos de mayor tamaño como coches. (*false*).
- *Seguridad*. Esta variable ha sido ya comentada, ya que es compartida con las oficinas. Informa de si el recinto, que alberga la zona de estacionamiento, está vigilado por seguridad privada (*true*) o no (*false*).

Puede observarse en la Tabla *24* el listado de campos específicos de cada uno de los tipos de inmuebles. Como puede apreciarse, las viviendas son los inmuebles que contienen un mayor número de campos adicionales, y las cocheras los que menor. Además, hay campos compartidos por más de un tipo de inmueble como puede ser la disponibilidad o no de aire acondicionado.

Tabla 24. *Tabla-resumen de los campos adicionales de cada uno de los proyectos.*

Viviendas	**Locales**	**Oficinas**	**Cocheras**
Cochera	Localización	Agua Caliente Indep	Puerta Autom
Ascensor	Esquina	Calefacción Indep	Parking Motoc
Piscina	Salida de humos	Seguridad	Seguridad
AireAcon	AireAcon	Distribución	
Terraza		Uso	
Trastero		Cochera	
Tendederos		Ascensor	
Empotrados		AireAcon	
Condición			
Estado			

Fuente: Elaboración propia a partir de los datos recogidos por Idealista

6.5. Instalación del programa

El software se distribuye a través de un archivo ejecutable, con 70 Mb de tamaño en disco aproximadamente, que se compone de un autoinstalable que deberá completar el usuario antes de comenzar su uso. Este instalador es, únicamente, compatible con el sistema operativo Windows, debido a la librería rJava, que, al necesitar definir variables de entorno en el sistema, limita el uso multiplataforma. No obstante, pueden crearse instaladores específicos del programa para otros sistemas operativos, ya que como se ha mencionado anteriormente, éste es compatible con diversas plataformas.

Como se ha indicado previamente, es importante recordar en este punto, que antes de iniciar la instalación del programa, el usuario debe tener instalada una versión de Java 1.8.0_40 o superior, el software R, y en éste, la librería rJava, que puede encontrarse en los repositorios habituales.

Para iniciar el proceso de instalación basta hacer doble clic en el archivo setup.exe. Al hacerlo, se abrirá el cuadro de diálogo que se muestra en la Figura 55.

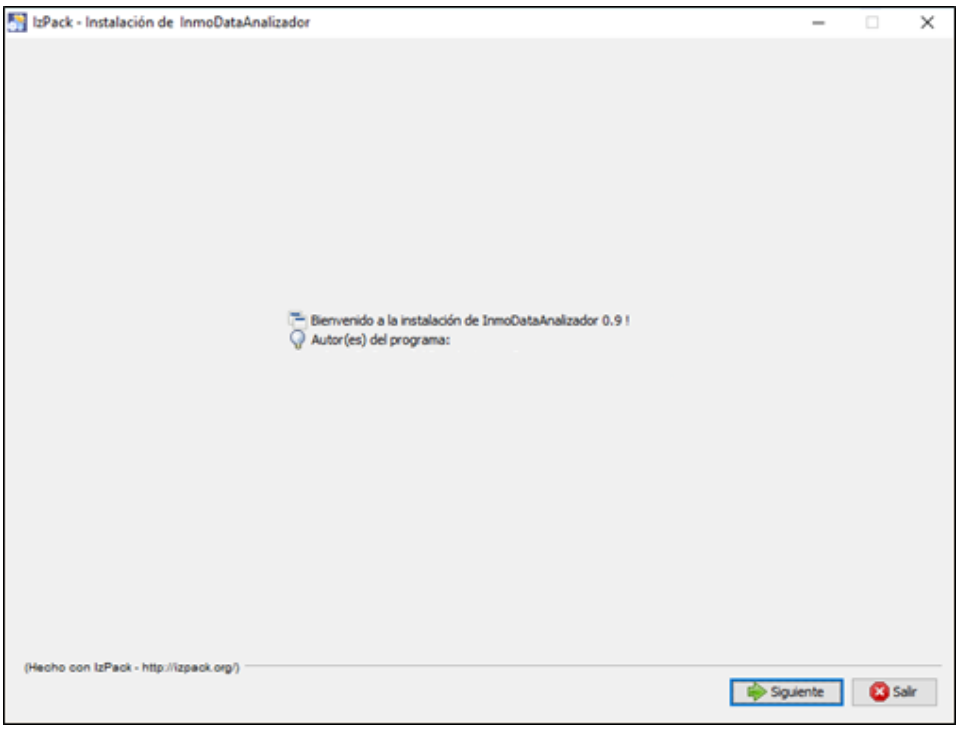


Figura 55. Ventana inicial de instalación.

Para comenzar la configuración de la instalación, basta con hacer clic en el botón *Siguiente*. En la siguiente ventana, Figura 56, que se muestra, el usuario deberá

seleccionar la carpeta donde desea que se sitúen los archivos del programa. Por defecto, esta carpeta será *c:\Program Files (x86)\InmoDataAnalizador*. Si se desea elegir la carpeta propuesta, basta con hacer clic, nuevamente, en el botón Siguiente. Si, por el contrario, se desea cambiar el destino de la instalación deberá hacer clic en el botón *Escoger…*, y elegir la carpeta deseada.

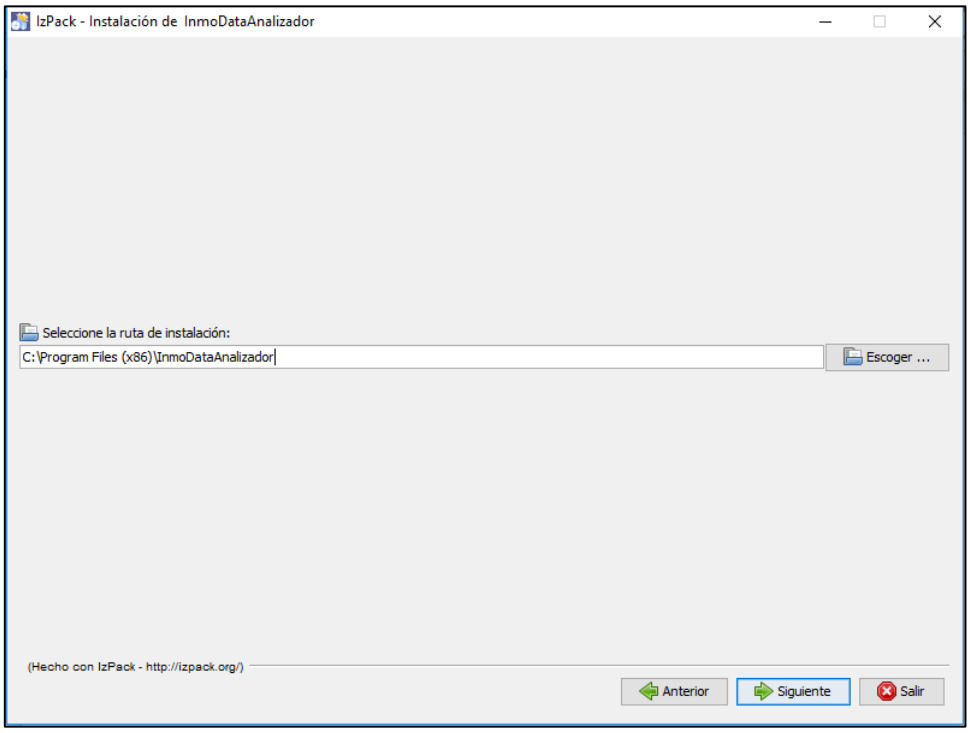


Figura 56. Instalación. Ventana de selección de destino

En la siguiente ventana se informa de los paquetes que se van a instalar, que son: el ejecutable con el programa en sí, las librerías Java, necesarias para su correcto funcionamiento, y los archivos R que contienen tanto las funciones R creadas para la realización de los estudios estadísticos, como los paquetes de los que tienen dependencias estas funciones. Como puede observarse en la Figura 57, desde esta ventana podrían seleccionarse, además, los elementos a instalar, entre los no obligatorios para el correcto funcionamiento del programa. No obstante, todos son necesarios, por lo que no es posible marcar o desmarcar alguna de las opciones. Para continuar a la siguiente ventana, basta con hacer clic, nuevamente, en el botón *Siguiente*.

El programa

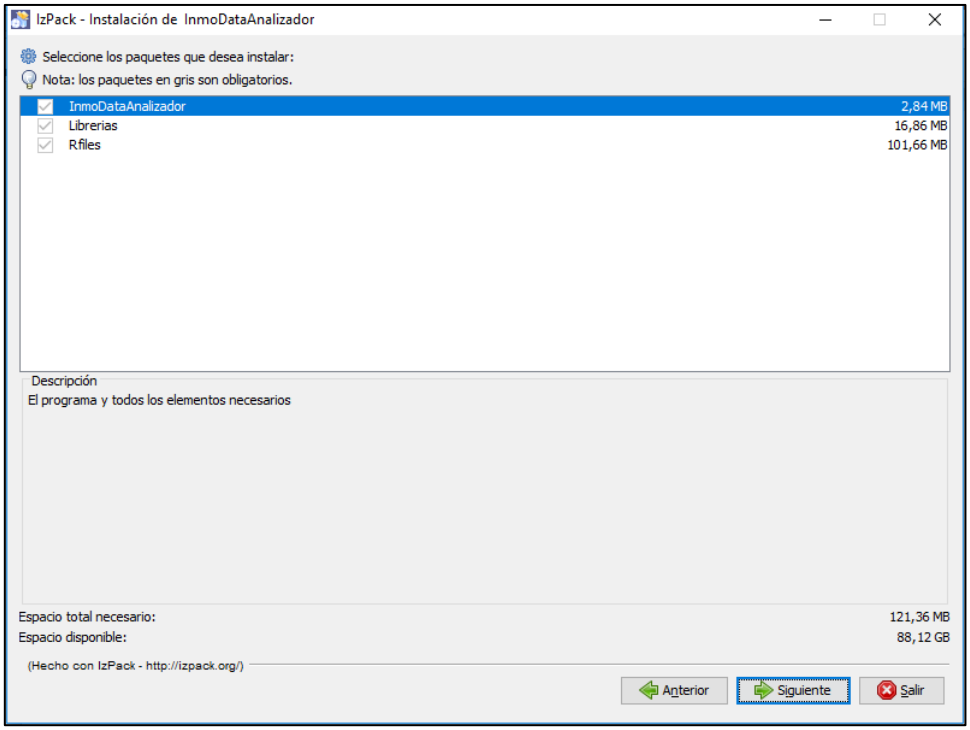


Figura 57. Instalación. Ventana de selección de paquetes

En la siguiente ventana, es donde tiene lugar la instalación del programa en el sistema. Este proceso se muestra a través de dos barras de progreso en las que se muestra el avance de la instalación de cada uno de los paquetes, así como el de la instalación total. Puede verse en la Figura 58.

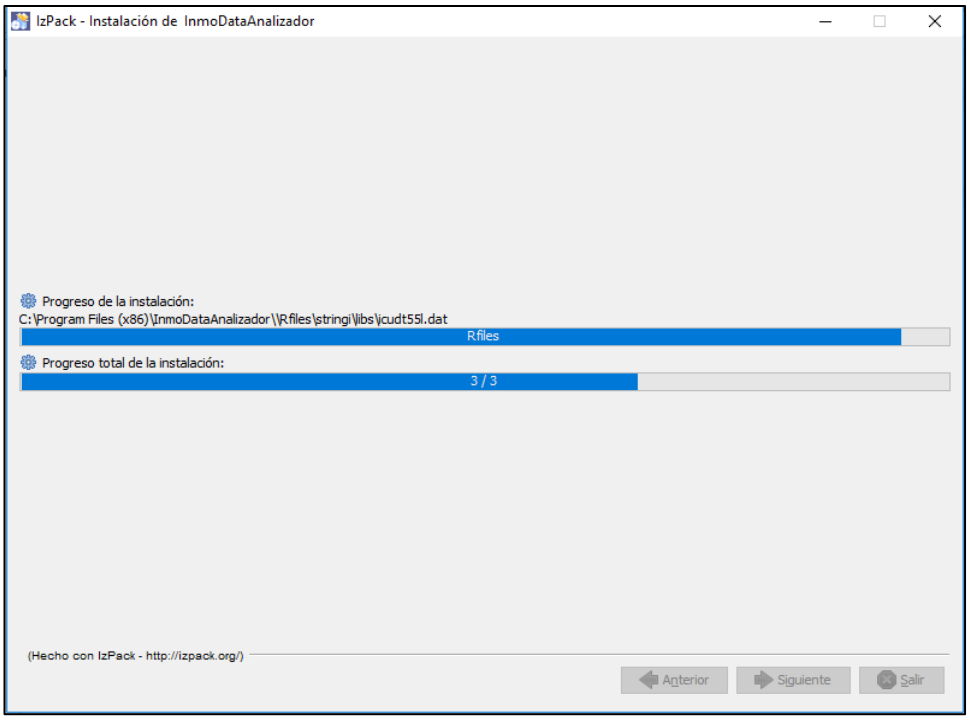


Figura 58. Instalación. Ventana de progreso de instalación

Al finalizar la instalación, un mensaje alertará de esta situación y bastará con pulsar sobre el botón *Siguiente* para continuar la configuración del programa.

Para el funcionamiento del enlace entre Java y el motor de R, es necesario determinar la localización exacta de dos bibliotecas de enlace dinámico o archivos dll, y generar con éstas, dos variables en el registro del sistema. Esto se hace a través de la siguiente ventana del instalador, mostrada en la Figura 59. En ella, el usuario deberá seleccionar la localización exacta de dos archivos creados previamente al inicio de la instalación, concretamente en la instalación de R y posteriormente, en la del paquete rJava.

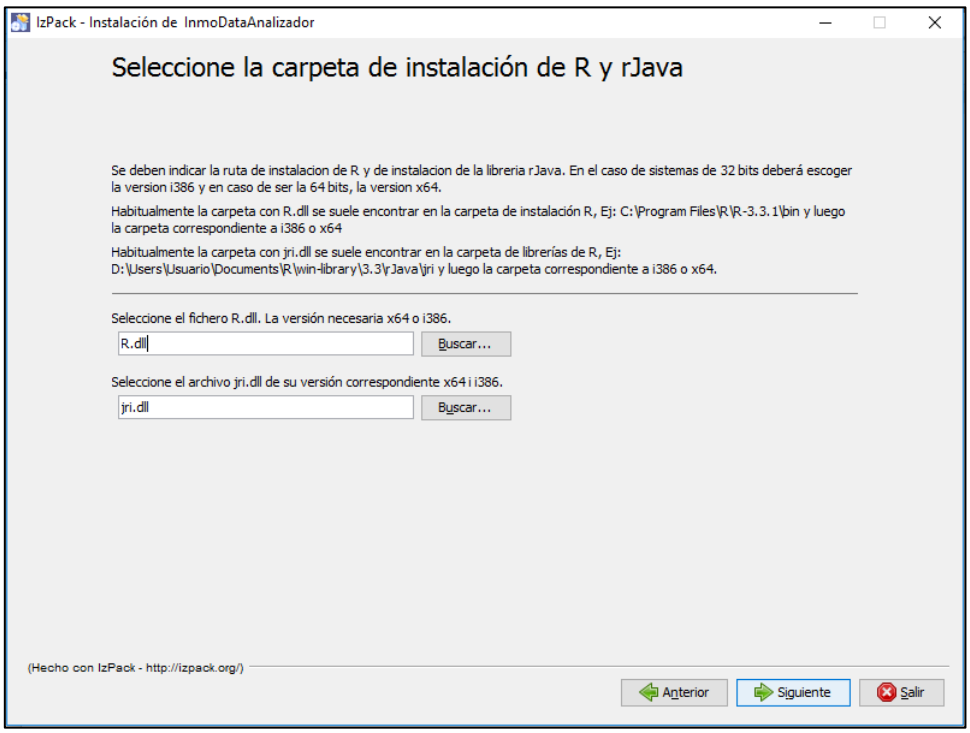


Figura 59. Instalación. Ventana de creación de variables de entorno

En la ventana se describe la localización usual de los archivos para los que solicita la ruta, pero esta localización dependerá de la versión de R disponible, de su configuración en la instalación, y de otras configuraciones del sistema.

Para indicar la localización de estos dos archivos, se deberá hacer clic en el botón *Buscar…* de cada uno de ellos, y a continuación se seleccionará el archivo requerido a través de la ventana de exploración, que se muestra en la Figura 60. Siguiendo las recomendaciones dadas, no será difícil encontrar la ubicación de estos archivos.

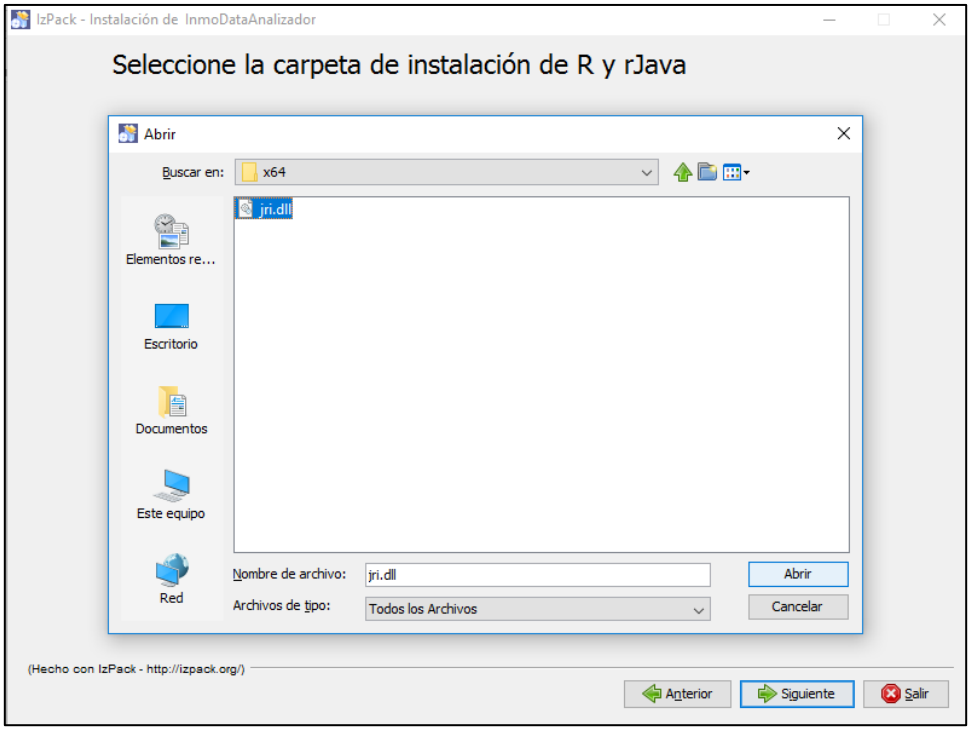


Figura 60. Instalación. Ventana de exploración

Una vez seleccionadas las ubicaciones de los archivos, basta con hacer clic en el botón Siguiente para que las variables de entorno sean generadas en el registro del sistema y pueda llevarse a cabo la conexión entre el programa y el motor de R.

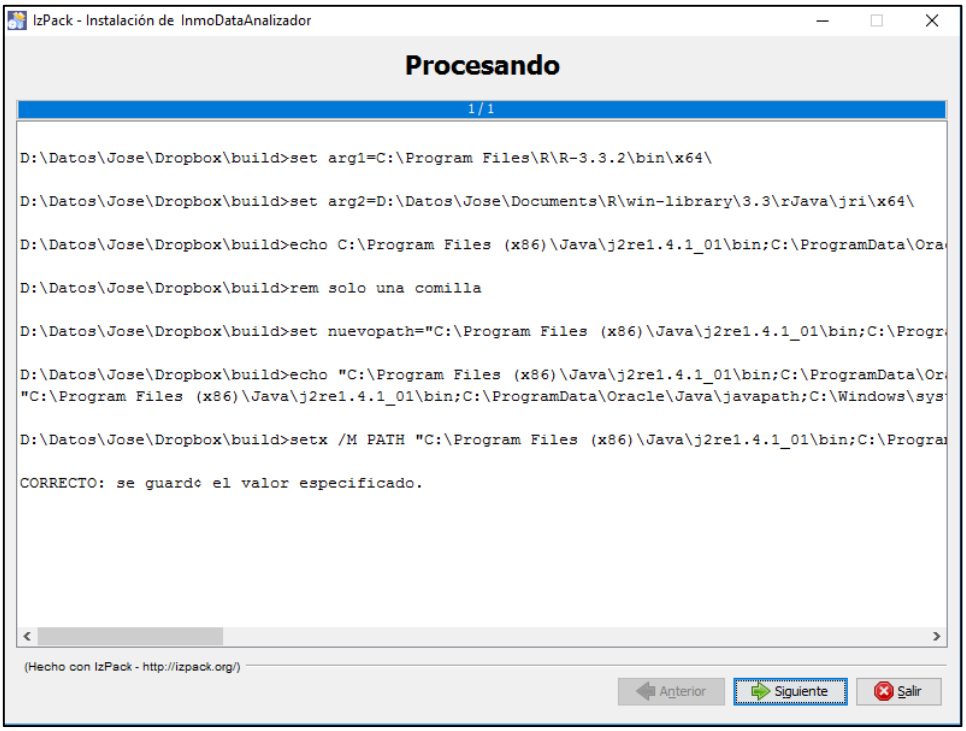


Figura 61.Instalación. Ventana de progreso de creación de variables de entorno

El progreso de creación de las estas variables y su resultado se muestra en la ventana mostrada en la Figura 61.

Tras comprobar que el proceso se ha realizado correctamente, haremos clic en el botón *Siguiente* para acceder a la última ventana de configuración de la instalación, en la que el usuario puede elegir dónde crear el acceso al programa en el grupo de programas de Windows, el nombre del grupo, y si desea accesos directos al programa en el menú inicio o en el escritorio. También podrá elegir si desea que el programa sea visible para todos los usuarios del sistema operativo, o sólo para el usuario actual. Todas estas opciones pueden verse en la Figura 62.

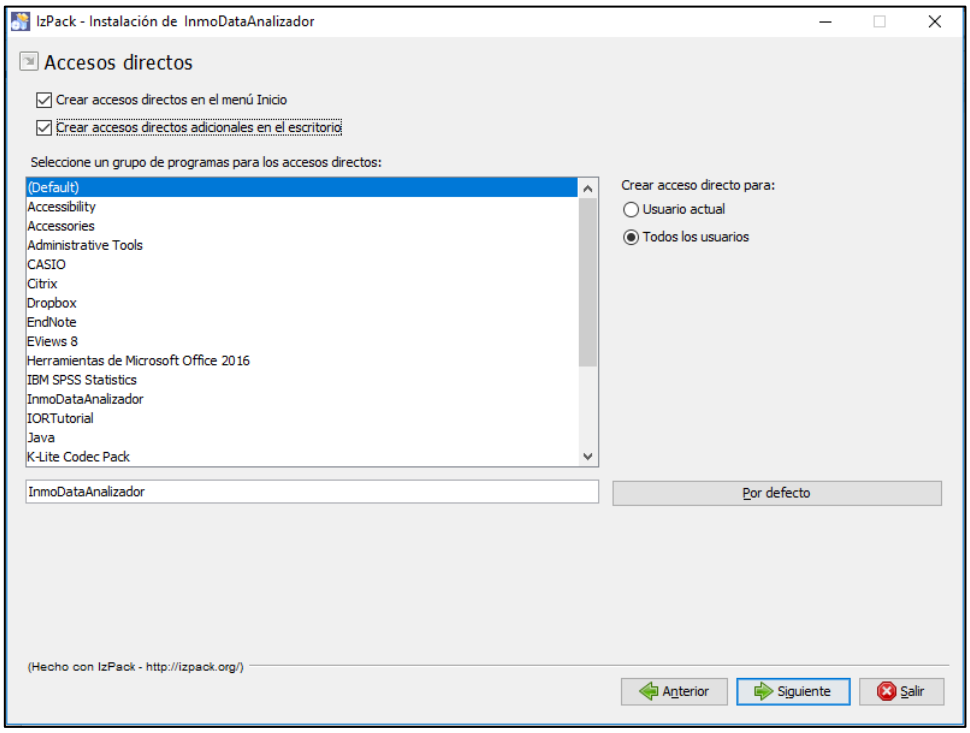


Figura 62. Instalación. Ventana de creación de accesos directos

Para finalizar la instalación, el usuario debe hacer clic, una vez más, sobre el botón *Siguiente*, para acceder a la ventana de cierre de la instalación. En ésta, se muestra un mensaje indicando que la instalación se ha realizado correctamente, e indicando, a su vez, que se ha generado un archivo para la correcta desinstalación del software, así como su ruta de acceso. Esta última venta del instalador del programa se muestra en la Figura 63. Para finalizar, basta con hacer clic en el botón *Hecho*.

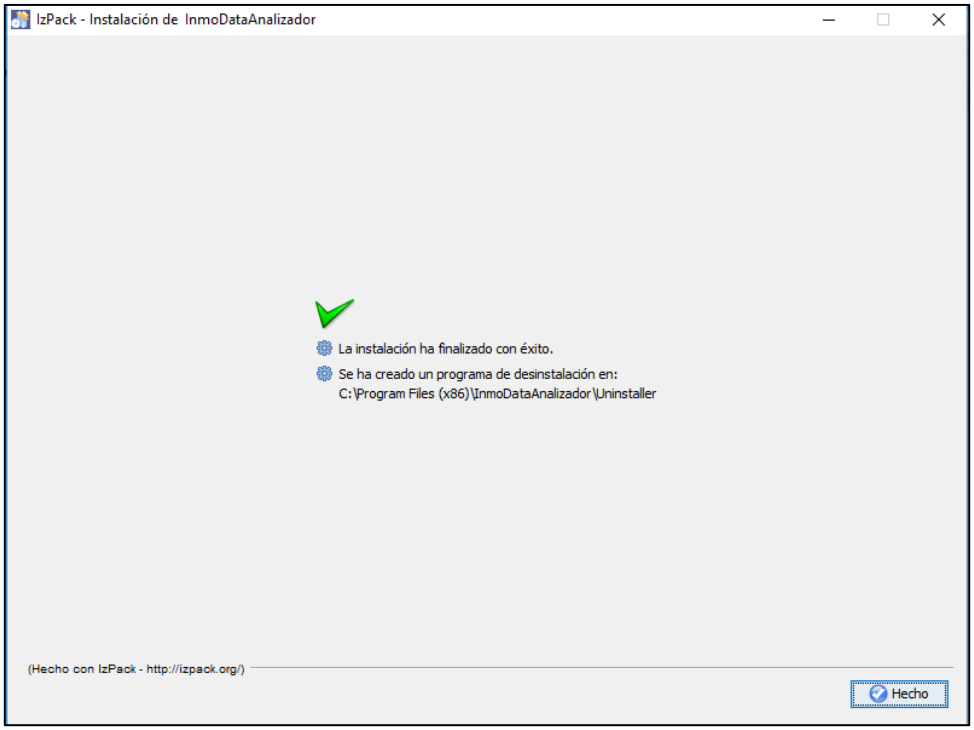


Figura 63. Instalación. Ventana de cierre

6.6. Funcionamiento del programa. Manual de usuario

Una vez instalado el programa correctamente, al hacer clic sobre el ejecutable, se inicia el menú principal del programa, que puede observarse en la Figura 64. Desde él se puede acceder a las principales funcionalidades del programa.

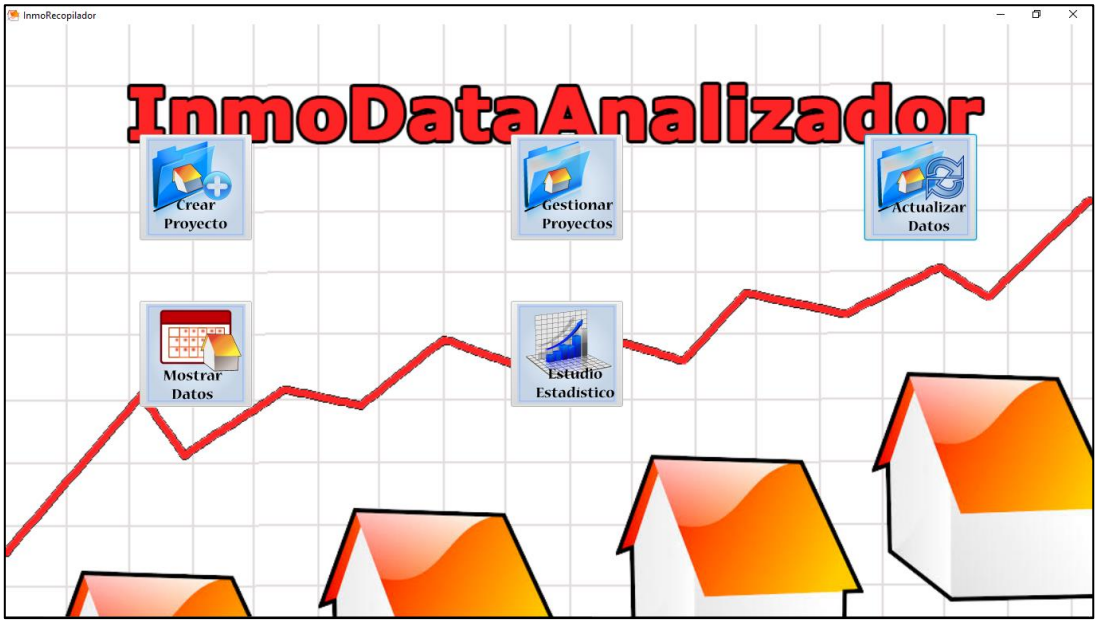


Figura 64. Ventana principal

Para acceder a ellas, basta con hacer clic sobre el botón correspondiente. Estos dan acceso a las siguientes ventanas:

- *Creación de proyectos*: Esta opción permite al usuario la creación de un nuevo proyecto mediante la búsqueda de inmuebles a través del servidor de idealista.com, así como el guardado de éste.
- *Gestión de proyectos*: Esta ventana está destinada a la consulta de proyectos guardados con anterioridad, y la selección del proyecto con el que se desea trabajar en la sesión actual.
- *Actualización de datos*. Esta ventana permite al usuario actualizar los datos de los inmuebles contenidos en un proyecto creado con anterioridad.
- *Consulta, filtrado y exportación de datos*: Una vez seleccionado el proyecto con el que se va a trabajar, el botón mostrar datos enlaza con la ventana desde la que se pueden visualizar las características de los inmuebles contenidos en él, pudiendo ordenar o filtrar los datos según cualquier variable analizada o exportarlos a formato CSV.
- *Análisis estadístico de los datos*: Esta opción permitirá al usuario realizar un estudio estadístico completo de los datos del proyecto seleccionado.

En un primer momento, sólo los dos primeros botones darán acceso a las ventanas correspondientes, ya que el resto de opciones requiere para su ejecución que se haya seleccionado previamente un proyecto. En caso de que el usuario haga clic en alguno de ellos, el programa indicará esta imposibilidad y abrirá automáticamente la ventana de gestión de proyectos.

Analizaremos, a continuación, en profundidad, cada una de las ventanas descritas anteriormente.

6.6.1. Ventana de creación de proyectos

Esta ventana permite al usuario realizar una búsqueda de inmuebles de la base de datos del portal idealista.com. La Figura 65 muestra esta ventana.

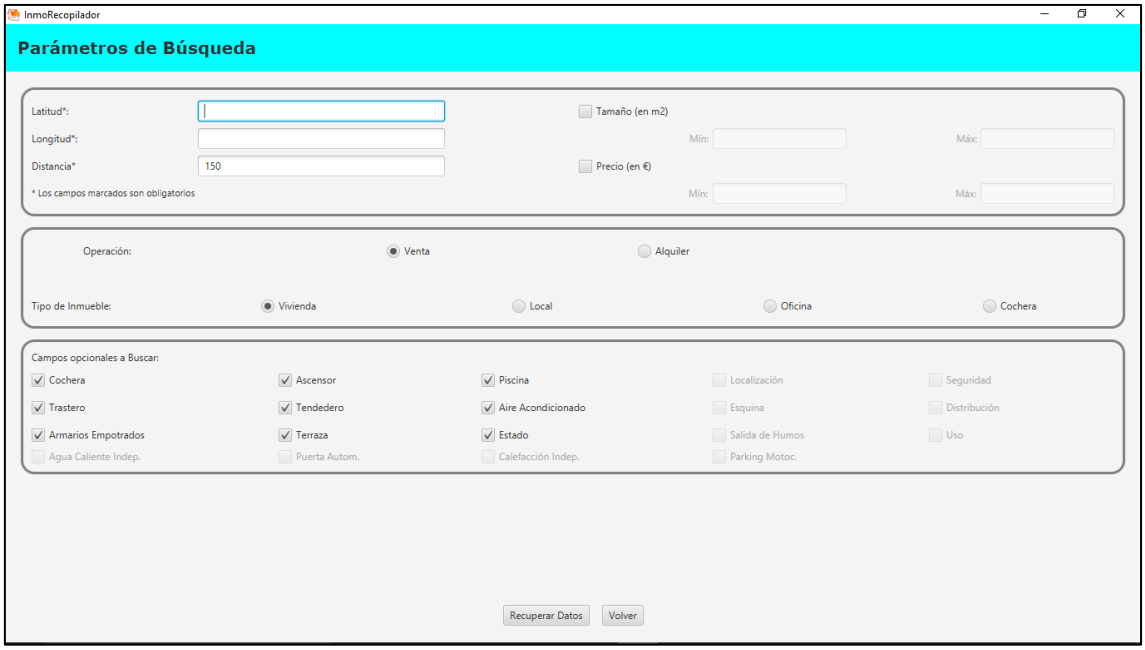


Figura 65. Ventana de creación de proyectos

Tras seleccionar las opciones deseadas, el usuario, a través del botón *recuperar datos* realizará una conexión al servidor del portal inmobiliario idealista, a través de la que llevará a cabo la consulta de inmuebles disponibles que verifiquen dichas condiciones. Es obvio que para el establecimiento de esta conexión es necesario que el usuario disponga de conexión a internet en el momento de la solicitud. Como puede observarse, esta ventana esta dividida en tres bloques:

- En el bloque superior se indicará la latitud y la longitud del punto geográfico que ocupará el centro de la búsqueda. Este punto deberá pertenecer a cualquier localización del estado español. El cuadro de texto *Distancia* indica máximo radio, en metros, desde el punto central, en el que se realizará la búsqueda de inmuebles. En esta sección además, puede realizarse un filtrado de los inmuebles a buscar por precio (en euros) y por tamaño (en metros cuadrados). Este último filtro no está disponible para la búsqueda de cocheras, debido a que no se dispone de esta información en este tipo de inmuebles.
- El bloque central permite seleccionar el tipo de inmueble y de operación que se desea analizar. Se puede seleccionar entre vivienda, local, oficina y cochera; y en todos ellos, a su vez, entre venta y alquiler.
- La búsqueda básica incluye una serie de campos que depende del tipo de inmueble seleccionado, pero además también es posible seleccionar entre un conjunto de campos adicionales para ampliar esta información. Estos campos pueden marcarse en el tercer bloque de la ventana de creación de proyectos. Estos campos son, en función del tipo de inmueble, los siguientes:
    - Vivienda (*home*): *cochera, ascensor, piscina, trastero, tendedero, aire acondicionado, armarios empotrados*, y *terraza*; con la información de si dispone o no de cada uno de ellos. También está disponible el campo *Estado*

con tres valores posibles: *good*, *renew* y *newdevelopment*, es decir, si el inmueble está en buen estado, reformado o es de nueva construcción.

- Local (*premise*): *Aire acondicionado* y *salida de humos*; con la información de si dispone de ellos, el campo *esquina*, que indica si el local está en la intersección de dos calles, y el campo *localización* que devuelve información sobre si el local se encuentra a pie de calle (*street*), si forma parte de un centro comercial (*shoppingcenter*), o si está situado en una entreplanta (*mezzanine*).
- Oficina (*office*): *Cochera*, *Agua caliente independiente*, Ascensor, *Aire acondicionado*, *Calefacción independiente* y *Seguridad;* con valores lógicos que indican si el inmueble dispone o no de estas características. El campo *Distribución* devuelve la información sobre si la oficina es diáfana (*open*) o si por el contrario tiene una distribución dividida en estancias o tabicada (*close*). Por último, disponemos del campo *Uso* que hace referencia al edificio en el que se encuentra la oficina. Éste puede estar destinado en exclusiva a oficinas (*exclusive*) o por el contrario, puede contener también viviendas (*mixed*).
- Cochera (*garage*): *Seguridad* y *puerta automática*; con información sobre su disponibilidad. También dispone del campo *parking de motocicletas*, para el caso de que la cochera esté destinada a este tipo de vehículos por sus dimensiones.

Una vez seleccionadas las opciones deseadas de la ventana de creación de proyecto, basta con pulsar el botón recuperar datos para que se inicie la búsqueda. Tras la finalización de ésta, se muestra una ventana de previsualización, Figura 66, de los inmuebles que verifican estas opciones. En ella se muestran los primeros 50 resultados de la búsqueda y se indica el total de resultados obtenidos.

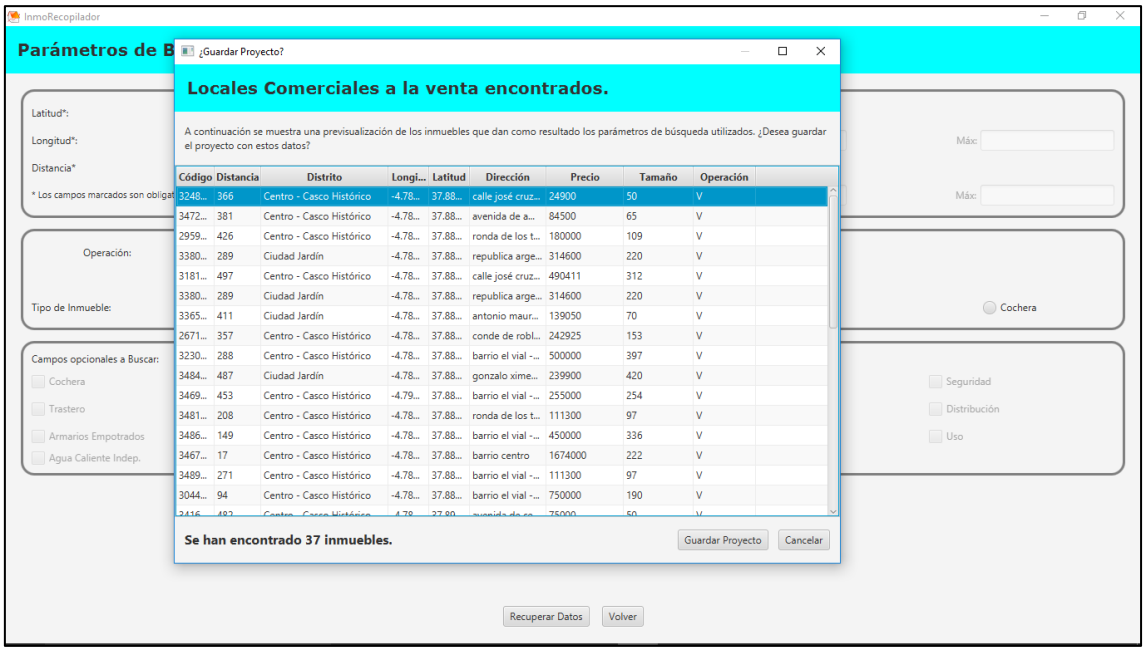


Figura 66. Ventana de vista previa de resultados de la búsqueda

En este momento, el usuario podrá guardar el proyecto o cancelarlo. Si se selecciona la primera de las opciones, se abrirá un cuadro de diálogo en el que se deberá indicar el nombre del proyecto. Tras aceptar, éste se integrará en el listado de la ventana de gestión de proyectos.

Es importante indicar que este proceso sólo genera el proyecto, definiendo las características del mismo, y lo incluye en el listado correspondiente. No obstante, y aunque el usuario ha podido observar en la ventana de previsualización las características de algunos inmuebles, ninguno de estos se han incluido hasta ahora en la base de datos. Por tanto, el proyecto creado hasta este momento está vacío de inmuebles.

6.6.2. Ventana de gestión de proyectos

Su función es la de gestionar los proyectos definidos anteriormente a través de la ventana de creación de proyectos. Como puede observarse en la Figura 67, en ella se muestra un listado de los proyectos existentes que puede ser ordenado desde cualquier campo tanto de forma ascendente como descendente.

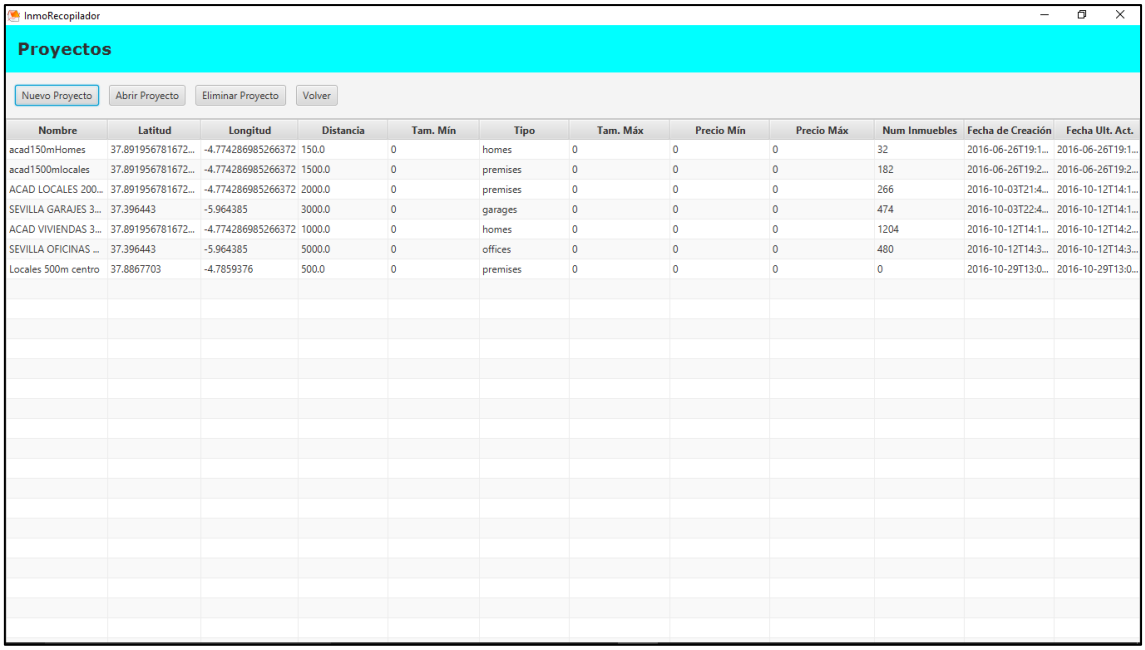


Figura 67. Ventana de gestión de proyectos

Los información disponible de cada proyecto es: *nombre* del proyecto, *latitud* y *longitud* del centro de la búsqueda, *distancia* o radio de búsqueda, el *tamaño mínimo* y

máximo por el que se ha realizado el filtrado, el *tipo* de inmueble, el *precio mínimo* y *máximo* para el que se ha filtrado, *el número de inmuebles* hallados en el proceso de búsqueda, la *fecha de creación* del proyecto y la *fecha de la última actualización* llevada a cabo. Como puede observarse en la captura mostrada en la Figura 67, el último proyecto, que se corresponde con la búsqueda mostrada en la Figura 66, no contiene ningún inmueble. Esto es debido a que al crearlo, como se ha mencionado anteriormente, no se ha guardado la información de ningún inmueble, sólo los resultados de la búsqueda solicitada.

En la parte superior de la ventana de gestión de proyectos, pueden observarse cuatro botones desde los que el usuario podrá:

- Volver a la ventana de creación de proyectos pulsando en el botón *nuevo proyecto*.
- Seleccionar uno de los proyectos del listado. Para ello deberá seleccionar uno de los proyectos y pulsar el botón *abrir proyecto*. También puede seleccionarse haciendo doble clic sobre uno de los proyectos de la lista.
- Eliminar uno de los proyectos incluido en el listado. Basta con seleccionarlo de la lista y pulsar el botón *eliminar proyecto*. Un cuadro de diálogo se abrirá al intentar eliminar un proyecto solicitando confirmación de la acción, ya que una vez eliminado el proyecto se eliminarán todos los inmuebles contenidos en él, y el proceso no es reversible.
- Volver a la ventana principal pulsando el botón *volver*.

Todas las acciones relativas a la actualización de proyectos, la consulta, filtrado y exportación de datos, y al análisis estadístico requieren que un proyecto esté activo, por lo que si el usuario intenta acceder a éstas, el programa le redirigirá a la ventana de gestión de proyectos.

6.6.3. Ventana de actualización de datos

Una vez creado un proyecto, para obtener los datos de los inmuebles que cumplen las condiciones de la búsqueda solicitada, el usuario tendrá que drigirse a la ventana de actualización de datos, donde se guardarán los datos de los inmuebles por primera vez. En la Figura 68 se muestra la actualización de datos de un proyecto concreto.

En esta actualización se obtiene tanto la información de los campos incluidos en la búsqueda básica como la de los campos opcionales de búsqueda que se han seleccionado

previamente en la ventana de creación de proyecto. Además se crea una variable nueva denominada *PrecioFecha* en el que la fecha se corresponde con el día en el que se ha realizado la actualización de los datos.

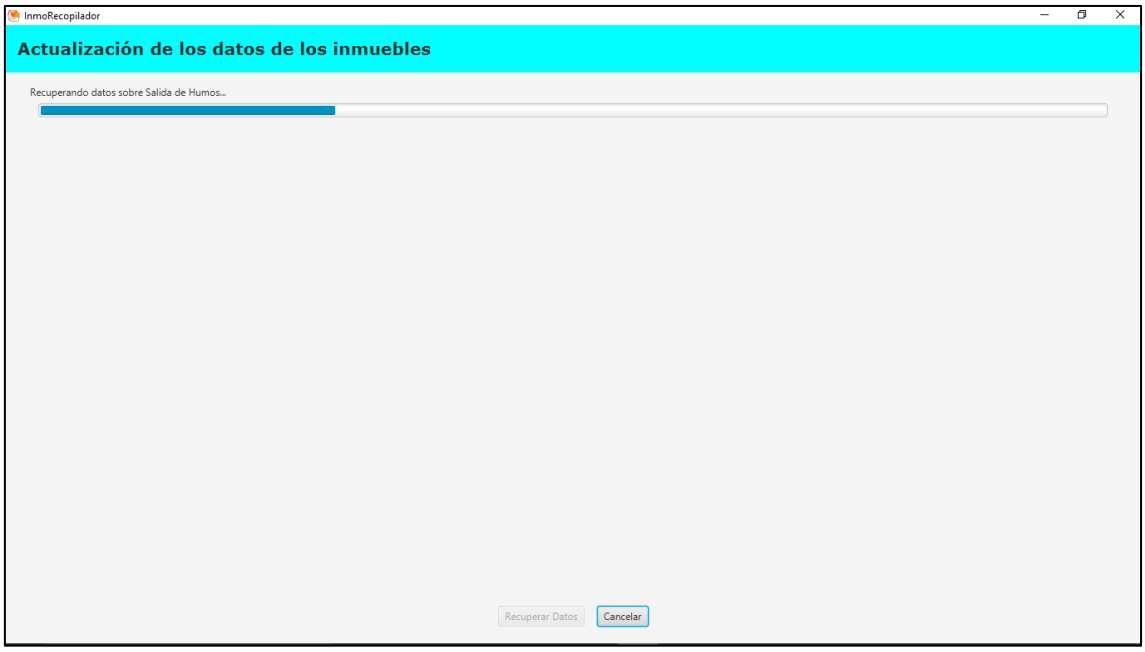


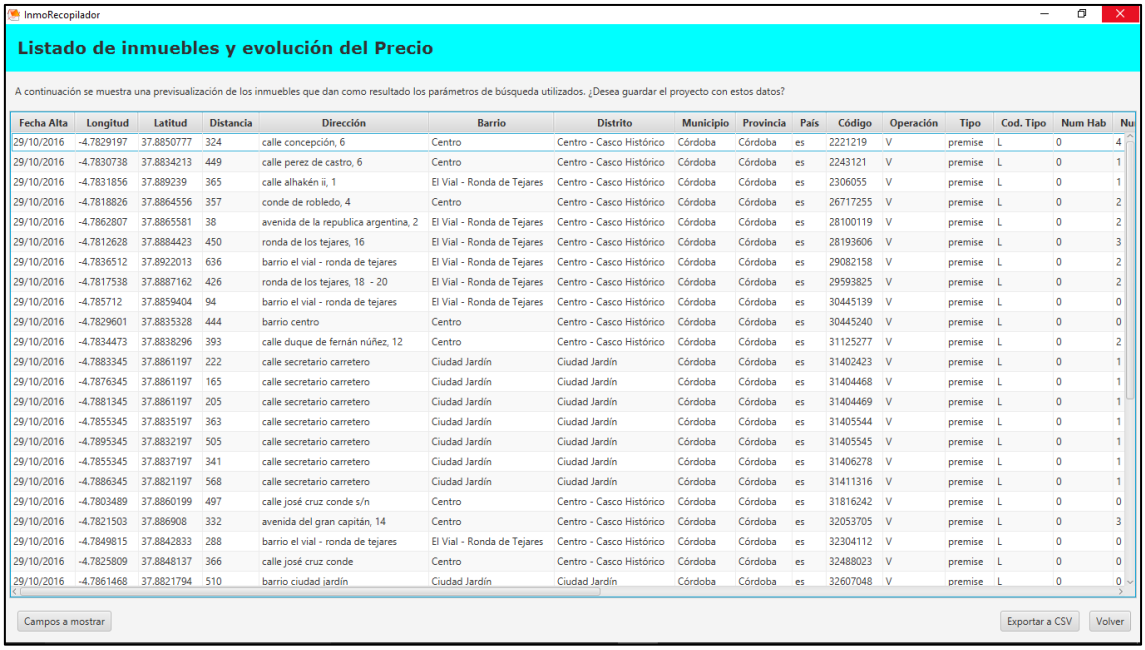
Figura 68. Ventana de actualización de datos

Una vez creado el proyecto y actualizados los datos, éste ya contiene la información de los inmuebles solicitados. Esta información podrá ser actualizada tantas veces como se desee en fechas posteriores, con el objetivo de crear una base de datos con información temporal. Cada una de estas actualizaciones creará una nueva variable con el nombre *PrecioFecha*, con la fecha correspondiente a la de la actualización. Se indicará que un inmueble ha dejado de pertenecer a la base de datos, asignando un valor perdido al valor de éste en el precio actualizado en esa fecha.

Los nuevos inmuebles se identifican por tener asignados valores perdidos en todos los precios salvo en el correspondiente a la última actualización. Para éstos, se obtiene también toda la información relativa a sus características, tanto básicas como opcionales.

6.6.4. Ventana de consulta, filtrado y exportación de datos

Una vez generado y actualizado un proyecto, se realizará la consulta de los resultados obtenidos a través de esta ventana, que se muestra en la Figura 69.



Fecha Alta	Longitud	Latitud	Distancia	Dirección	Barrio	Distrito	Municipio	Provincia	País	Código	Operación	Tipo	Cod. Tipo	Num Hab	Nu
29/10/2016	-4.7829197	37.8850777	324	calle concepción, 6	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	2221219	V	premise	L	0	4
29/10/2016	-4.7830738	37.8834213	449	calle perez de castro, 6	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	2243121	V	premise	L	0	1
29/10/2016	-4.7831856	37.889239	365	calle alhakén ii, 1	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	2306055	V	premise	L	0	1
29/10/2016	-4.7818826	37.8864556	357	conde de robledo, 4	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	26717255	V	premise	L	0	2
29/10/2016	-4.7862807	37.8865581	38	avenida de la republica argentina, 2	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	28100119	V	premise	L	0	2
29/10/2016	-4.7812628	37.8884423	450	ronda de los tejares, 16	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	28193606	V	premise	L	0	3
29/10/2016	-4.7836512	37.8922013	636	barrio el vial - ronda de tejares	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	29082158	V	premise	L	0	2
29/10/2016	-4.7817538	37.8887162	426	ronda de los tejares, 18 - 20	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	29593825	V	premise	L	0	2
29/10/2016	-4.785712	37.8859404	94	barrio el vial - ronda de tejares	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	30445139	V	premise	L	0	0
29/10/2016	-4.7829601	37.8835328	444	barrio centro	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	30445240	V	premise	L	0	0
29/10/2016	-4.7834473	37.8838296	393	calle duque de fernán núñez, 12	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	31125277	V	premise	L	0	2
29/10/2016	-4.7883345	37.8861197	222	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31402423	V	premise	L	0	1
29/10/2016	-4.7876345	37.8861197	165	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31404468	V	premise	L	0	1
29/10/2016	-4.7881345	37.8861197	205	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31404469	V	premise	L	0	1
29/10/2016	-4.7855345	37.8835197	363	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31405544	V	premise	L	0	1
29/10/2016	-4.7895345	37.8832197	505	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31405545	V	premise	L	0	1
29/10/2016	-4.7855345	37.8837197	341	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31406278	V	premise	L	0	1
29/10/2016	-4.7886345	37.8821197	568	calle secretario carretero	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	31411316	V	premise	L	0	1
29/10/2016	-4.7803489	37.8860199	497	calle josé cruz conde s/n	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	31816242	V	premise	L	0	0
29/10/2016	-4.7821503	37.886908	332	avenida del gran capitán, 14	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	32053705	V	premise	L	0	3
29/10/2016	-4.7849815	37.8842833	288	barrio el vial - ronda de tejares	El Vial - Ronda de Tejares	Centro - Casco Histórico	Córdoba	Córdoba	es	32304112	V	premise	L	0	0
29/10/2016	-4.7825809	37.8848137	366	calle josé cruz conde	Centro	Centro - Casco Histórico	Córdoba	Córdoba	es	32488023	V	premise	L	0	0
29/10/2016	-4.7861468	37.8821794	510	barrio ciudad jardín	Ciudad Jardín	Ciudad Jardín	Córdoba	Córdoba	es	32607048	V	premise	L	0	0

Figura 69. Ventana de consulta, filtrado y exportación de datos

Buena parte del espacio de ventana lo ocupa la tabla de datos del proyecto activo. Como se ha comentado anteriormente, los campos que se muestran dependen del tipo de inmueble contenido en el proyecto.

Al hacer doble clic sobre la información de uno de los inmuebles, se abre una ventana del navegador con la información del inmueble en la propia web del portal *idealista.com*.

En la parte inferior izquierda de la pantalla se sitúa el botón de *campos a mostrar* en la visualización de los datos. Si hacemos clic sobre él, se abrirá un cuadro de diálogo como el que se muestra en la Figura 70.

Algunos de los campos disponibles están deshabilitados por defecto debido a que el servidor de Idealista no provee información alguna sobre los mismos. La decisión sobre dejar la opción de habilitarlos por parte del usuario se sustenta en la posibilidad de que en un futuro, el proveedor de la información comience a proporcionarla.

Esta ventana está compuesta, a su vez, por cuatro pestañas con los campos que pueden ser mostrados en la ventana de consulta.

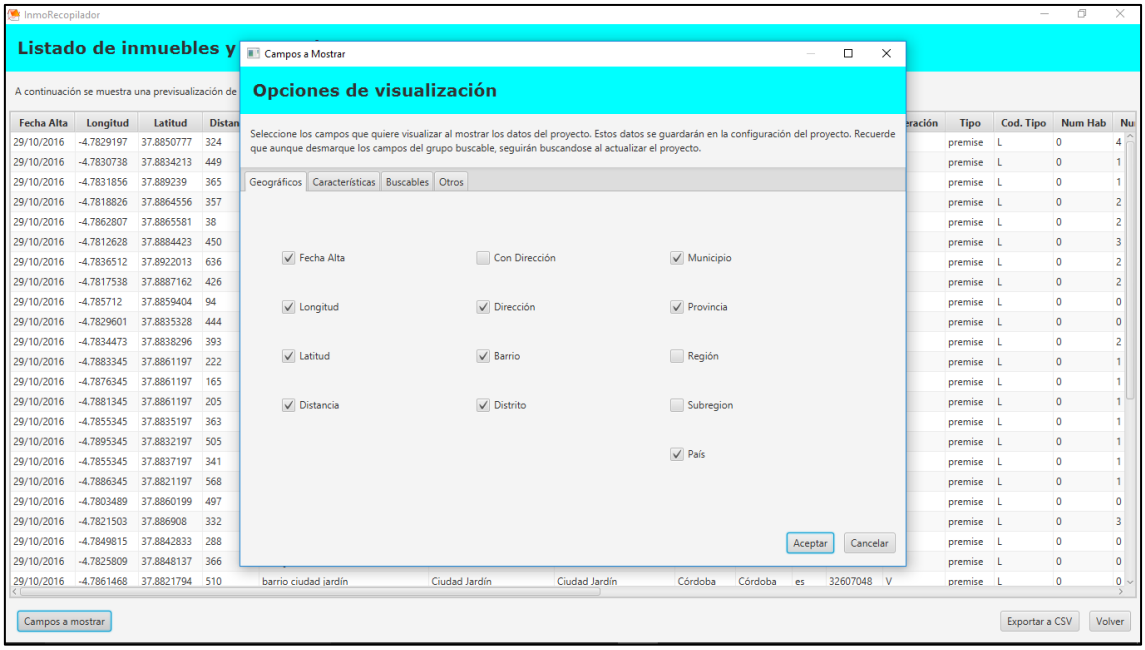


Figura 70. Opciones de visualización de la ventana de consulta. (Fuente: Elaboración propia)

Estos campos son:

- *Geográficos* (Figura 70):
  - Fecha Alta.
  - Con Dirección: campo deshabilitado por defecto ya que siempre toma el valor false.
  - Municipio.
  - Longitud.
  - Dirección.
  - Provincia.
  - Latitud.
  - Barrio.
  - Región: deshabilitado por defecto ya que siempre toma valores perdidos.
  - Distancia.
  - Distrito.
  - Subregión: deshabilitado también por no obtener resultados en este campo.
  - País.
- *Características* (Figura 71):
  - *Código* con el que se identifica el inmueble en la base de datos, único para cada uno.
  - *Descripción*: deshabilitado por defecto por no devolver información alguna.
  - *Piso*: localización en altura del inmueble.
  - *Operación*: Los valores que toma son venta: *V* y alquiler: *A*.

- *Número de habitaciones*.
- *Exterior*: deshabilitado por tomar siempre el valor *false*.
- *Tipo*: Identifica el tipo de inmueble.
- Número de baños.
- Condición: también deshabilitado por defecto por no ofrecer información alguna.
- *Código tipo*.
- Tamaño.
- Precio.

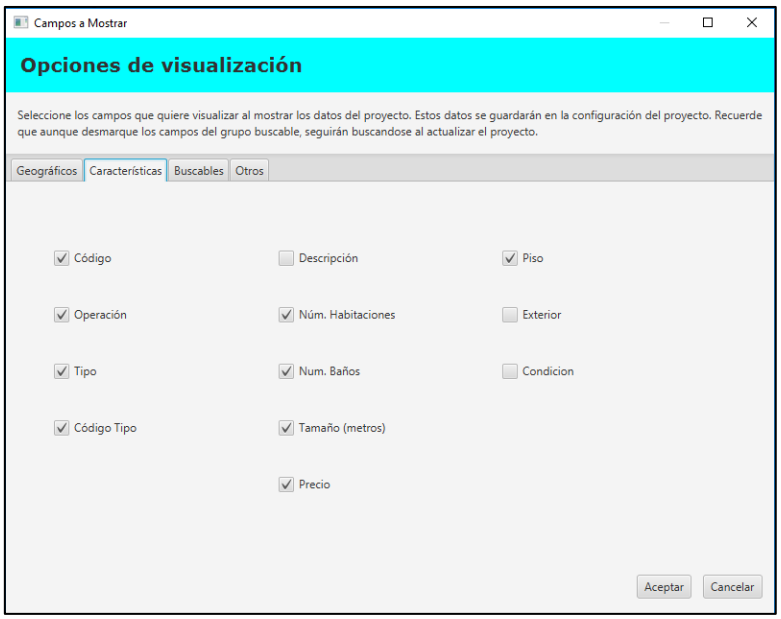


Figura 71. Opciones de visualización: Características

- *Buscables* (Figura 72):
  - *Cochera* (Disponible para viviendas y oficinas).
  - *Armarios empotrados* (Disponible para viviendas).
  - *Seguridad* (Disponible para cocheras y oficinas).
  - *Ascensor* (Disponible para viviendas y oficinas).
  - *Estado* (Disponible para viviendas).
  - *Distribución* (Disponible para oficinas).
  - *Piscina* (Disponible para viviendas).
  - *Localización* (Disponible para locales).
  - *Uso* (Disponible para oficinas).
  - *Aire acondicionado* (Disponible para viviendas, locales y oficinas).
  - *Esquina* (Disponible para locales).
  - *Puerta Automática* (Disponible para cocheras).

- *Terraza* (Disponible para viviendas).
- *Salida de humos* (Disponible para locales).
- *Parking Motocicletas* (Disponible para cocheras).
- *Trastero* (Disponible para viviendas).
- *Agua caliente independiente* (Disponible para oficinas).
- *Tendedero* (Disponible para viviendas).
- *Calefacción independiente* (Disponible para oficinas).

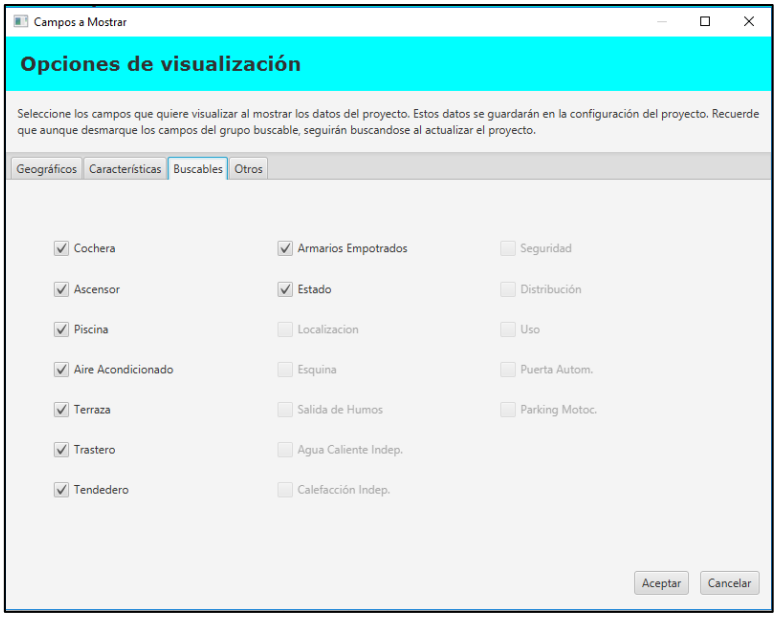


Figura 72. Opciones de visualización: Buscables.

- *Otros* (Figura 73):
  - Num Fotos.
  - Video. Siempre se devuelve el valor false. Por defecto está deshabilitado, por este motivo.
  - URL Fotos.
  - Tipo de Video: Siempre devuelve el valor F por lo que está deshabilitado por defecto.
  - URL.
  - Usuario. Este campo está deshabilitado por defecto debido a que no proporciona información alguna.
  - Miniaturas.
  - Tipo de Usuario. Este campo está deshabilitado por defecto debido a que no proporciona información alguna, siempre devuelve el valor 0.
  - URL thumbnails.
  - Comentario favorito: Este campo está deshabilitado por defecto debido a que no proporciona información alguna.

- Agencia. Está deshabilitada por defecto debido a que siempre devuelve el valor false.
- Favorito. Este campo está deshabilitado por defecto debido a que no proporciona información alguna.
- URL Logo Agencia. Este campo está deshabilitado por defecto debido a que no proporciona información alguna.

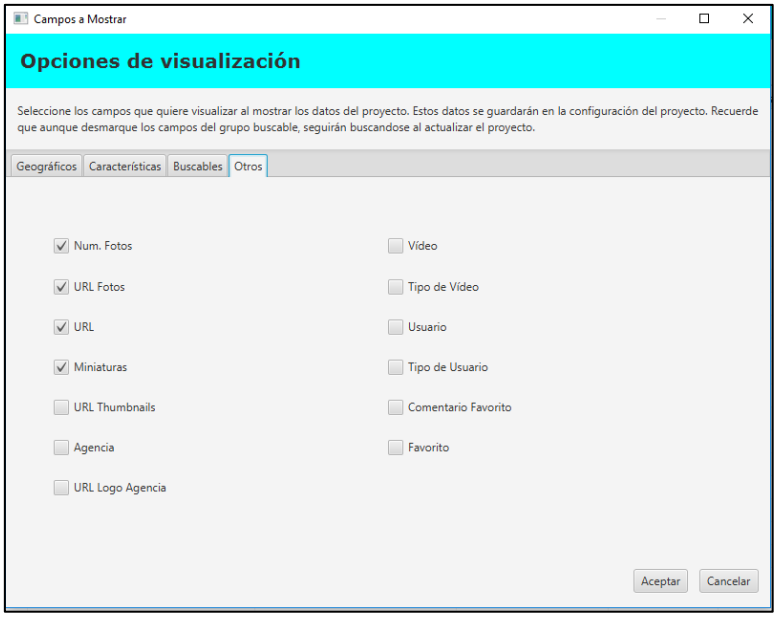


Figura 73. Opciones de visualización: Otros

Además de seleccionar los campos que se desean mostrar, desde esta ventana, el usuario podrá ordenar la tabla de datos a partir de cualquiera de los campos mostrados. Para ello, basta con hacer clic, con el botón izquierdo, en la celda en la que se muestra el nombre de la variable a partir de la que se desea realizar la ordenación. Una vez hecho esto, los inmuebles se ordenarán en sentido creciente, y esto se indicará con un triángulo a la derecha del nombre, como se observa en la Figura 74.

Hab	Num Baños	Tamaño	Precio ▲	Piso	Exterior	Co
	1	59	46000	2	false	fals
	1	70	48000	2	false	fals
	1	51	50000		false	fals
	1	60	51000	bj	false	fals
	1	60	51000		false	fals
	1	82	52000	2	false	fals
	1	82	52000	2	false	fals
	1	80	54000		false	fals
	1	70	56000	2	false	fals
	1	76	56000	2	false	fals
	1	60	56900	1	false	fals

Figura 74. Ordenación creciente.

Hab	Num Baños	Tamaño	Precio ▼	Piso	Exterior	Co
	5	300	850000	bj	false	tru
	5	790	850000		false	tru
	5	585	795000		false	tru
	4	300	775000		false	tru
	3	200	750000		false	fals
	5	300	750000	5	false	fals
	3	690	720000		false	fals
	2	450	700000		false	tru
	3	500	660666	6	false	tru
	4	234	650000		false	tru
	3	298	650000	5	false	tru

Figura 75. Ordenación decreciente.

Al hacer de nuevo clic, la ordenación será en sentido decreciente, y esto será indicado con el triángulo anterior invertido, como se muestra en la Figura 75. Al hacer un tercer clic, la ordenación es eliminada, volviendo a la ordenación inicial.

También es posible realizar filtrados a partir de los valores de las variables. Para ello, basta con, al contrario que para la ordenación, hacer clic con el botón derecho del ratón sobre el nombre de cualquiera de las variables. Esto abrirá un cuadro de diálogo como el que se muestra en la Figura 76, en la que se puede observar el menú de filtrado para la variable *Tipo* para un proyecto de viviendas.

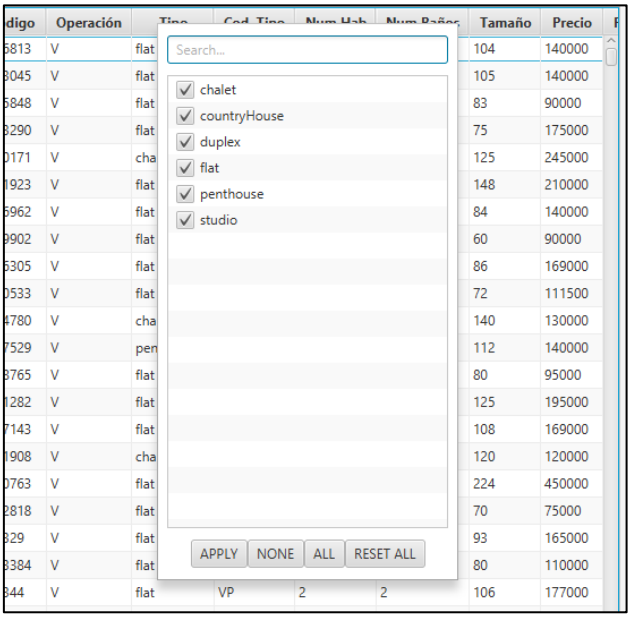


Figura 76. Menú filtrado de la ventana de consulta de datos

Como puede observarse, este cuadro de diálogo está dividido en tres bloques:

- El bloque superior está formado por una barra de búsqueda en el que el usuario podrá buscar las distintas modalidades de la variable sobre la que desea realizar el filtrado.
- El bloque central muestra todos los valores distintos que toma dicha variable. Esta lista es sensible al texto introducido en la barra de búsqueda del bloque superior, de modo que si, por ejemplo, escribimos la letra c en ésta, la lista se verá reducida únicamente a las modalidades que empiezan por c, en el caso de la Figura 76, *chalet* y *countryHouse*. A la izquierda de cada modalidad podemos observar la existencia de una casilla de verificación o checkbox, con la que se marcarán o desmarcarán según se desee incluir o excluir cada una de las modalidades en el filtro.
- El bloque inferior del cuadro de diálogo está compuesto por cuatro botones que componen el menú de filtrado: el botón *APPLY* con el que se ejecutará el filtrado

solicitado, el botón *NONE* desmarcará todas las casillas de verificación, el botón *ALL* que marcará todas las casillas de verificación y el botón *RESET ALL* que eliminará todos los filtros realizados.

El filtrado por una o más variables se mostrará con un icono con forma de embudo a la izquierda del nombre de la variable o variables para las que se ha llevado a cabo esta acción. En la Figura 77 puede observarse un ejemplo en el que se han realizado dos filtrados, uno para la variable municipio y otro para el tipo de inmueble. Además, se ha ordenado por el precio de éste en forma creciente.

Barrio	Distrito	Municipio	Provincia	País	Código	Operación	Tipo	Cod. Tipo	Num Hab	Num Baños	Tamaño	Precio	Pis
	Centro - Casco Histórico	Córdoba	Córdoba	es	33708189	V	flat	VP	3	1	116	7000	
	Levante - Lepanto - Fátima	Córdoba	Córdoba	es	34409140	V	flat	VP	3	1	69	40000	4
- Ollerías - Marrubial	Centro - Casco Histórico	Córdoba	Córdoba	es	34713613	V	flat	VP	2	1	55	42000	1
n José	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	34658320	V	flat	VP	3	1	59	46000	2
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	34648218	V	flat	VP	3	1	70	48000	2
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	32533272	V	flat	VP	2	1	51	50000	
- Ollerías - Marrubial	Centro - Casco Histórico	Córdoba	Córdoba	es	33152312	V	flat	VP	1	1	60	51000	bj
- Ollerías - Marrubial	Centro - Casco Histórico	Córdoba	Córdoba	es	33163638	V	flat	VP	2	1	60	51000	
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	33953492	V	flat	VP	3	1	82	52000	2
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	34555884	V	flat	VP	3	1	82	52000	2
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	32877065	V	flat	VP	2	1	80	54000	
umbacón - Camping	Santa Rosa - Valdeolleros	Córdoba	Córdoba	es	33745194	V	flat	VP	2	1	70	56000	2
	Levante - Lepanto - Fátima	Córdoba	Córdoba	es	33790452	V	flat	VP	2	1	76	56000	2
	Levante - Lepanto - Fátima	Córdoba	Córdoba	es	33803586	V	flat	VP	2	1	60	56900	1

Figura 77. Ejemplo de filtrado y ordenación de los datos

El menú de filtrado aplicado en esta ventana fue implementado por *Controlsfx*, como se ha mencionado previamente. Se ha mantenido el menú en el idioma original en el que fue creado por mantener la herramienta original creada por su autor.

Una vez aplicados los filtros deseados y con el orden escogido por el usuario, estos datos pueden ser exportados a formato *CSV* (*comma-separated values*), un formato simple de archivo en el que los datos de un mismo inmueble se separan por punto y coma. Las distintas filas, correspondientes a inmuebles, se separan por saltos de línea. El separador decimal es la coma.

Este formato de archivo es ampliamente aceptado por otros programas, fundamentalmente de tipo estadístico. Como ejemplos de software para el tratamiento de datos que tienen total compatibilidad con este formato, podemos destacar R, SPSS o Eviews.

Para realizar esta exportación de los datos, basta con hacer clic en el botón *Exportar a CSV* de la parte inferior derecha de la ventana. Esto abrirá un cuadro de diálogo como

el que se muestra en la Figura 78, en el que el usuario deberá seleccionar el directorio de destino del archivo y su nombre.

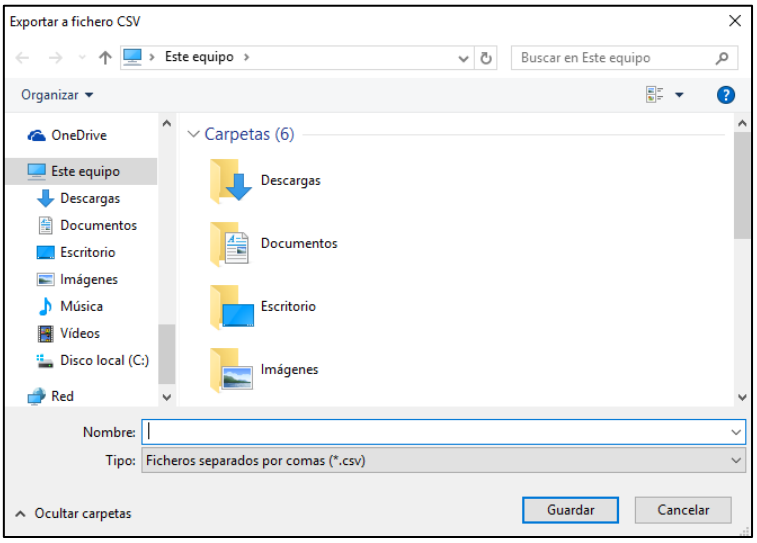


Figura 78. Cuadro de diálogo de guardado de archivo CSV

Es importante destacar que la selección de campos a mostrar, el filtrado y la ordenación de los datos sólo afecta a la visualización de los mismos y al archivo exportado a formato CSV. En ningún caso serán eliminados registros de inmuebles de la base de datos del proyecto.

6.6.5. Ventana de análisis estadístico de los datos

Desde esta ventana del programa, el usuario podrá realizar un análisis estadístico minucioso de los datos contenidos en el proyecto. Este análisis puede incluir un estudio descriptivo y gráfico de cada variable, análisis de asociación entre variables, construcción de modelos de regresión y validación de éstos para realizar estimaciones de la característica de un inmueble a partir del resto, determinación de la existencia de valores atípicos en alguna de las variables de estudio o en la construcción del modelo de regresión, estimación a través de una red neuronal del tipo perceptrón y su comparación con el modelo de regresión equivalente y representación en una imagen satélite de la distribución geográfica de los inmuebles con indicación de alguna característica concreta y su presencia en distintas zonas de la ciudad.

Todos los datos necesarios para la realización de los cálculos estadísticos son enviados a R, con el que el programa establece un puente de comunicación, quién realiza todo el proceso de cálculo y devuelve los resultados al programa, quien los imprime en pantalla en formato *HTML*.

Las variables disponibles para la realización de los estudios estadísticos dependen del tipo de inmueble que se está analizando. Estas no coinciden con las variables disponibles que ya han sido comentadas en la *ventana de consulta, filtrado y exportación de datos*; ya que de éstas se han eliminado aquellas que no tienen interés estadístico por ser únicas para cada inmueble como *Código* o por ser constantes como *País*. No obstante, las variables incluidas en este apartado pueden consultarse en el capítulo relativo a las funciones de R.

La ventana de análisis estadístico está compuesta por siete pestañas que se encuentran en la parte izquierda de ésta. También puede seleccionarse una pestaña haciendo clic sobre el botón circular situado en la parte inferior izquierda, como puede observarse en la Figura 79. Al hacerlo, emerge un menú en el que puede seleccionar la pestaña deseada.

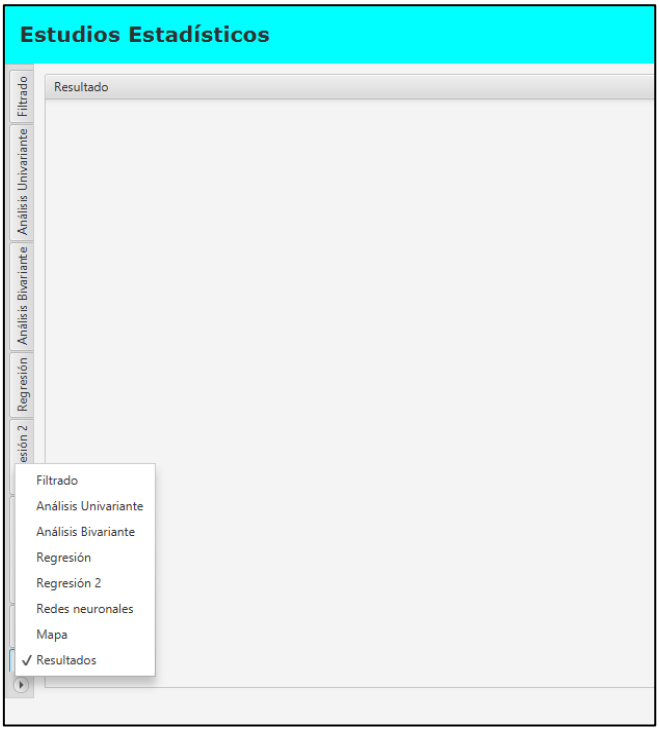


Figura 79. Pestañas de la ventana de Estudios Estadísticos

Las pestañas disponibles en esta ventana son:

- Filtrado.
- Análisis univariante.
- Análisis bivariante.
- Regresión.
- Regresión 2.
- Redes neuronales.
- Mapa.
- Resultado.

A continuación, se realizará una descripción detallada de cada una de éstas.

6.6.5.1. La pestaña Filtrado

La pestaña de filtrado es similar a la vista en la ventana de consulta de datos. Permite filtrar los datos de las variables que van a utilizarse en el estudio estadístico posterior. El funcionamiento es el mismo que el de la ventana de consulta. Basta con hacer clic con el botón derecho sobre el título de la variable, lo que desplegará el cuadro de diálogo mostrado en la Figura 76, y realizar el filtrado deseado.

El filtro realizado a través de esta pestaña sólo afectará a las observaciones incluidas en los estudios estadísticos realizados, por lo que la base de datos de inmuebles no se modificará al realizar filtros sobre la muestra dada.

La Figura 80 muestra esta pestaña con el menú de filtrado para la variable *Cochera* desplegado.

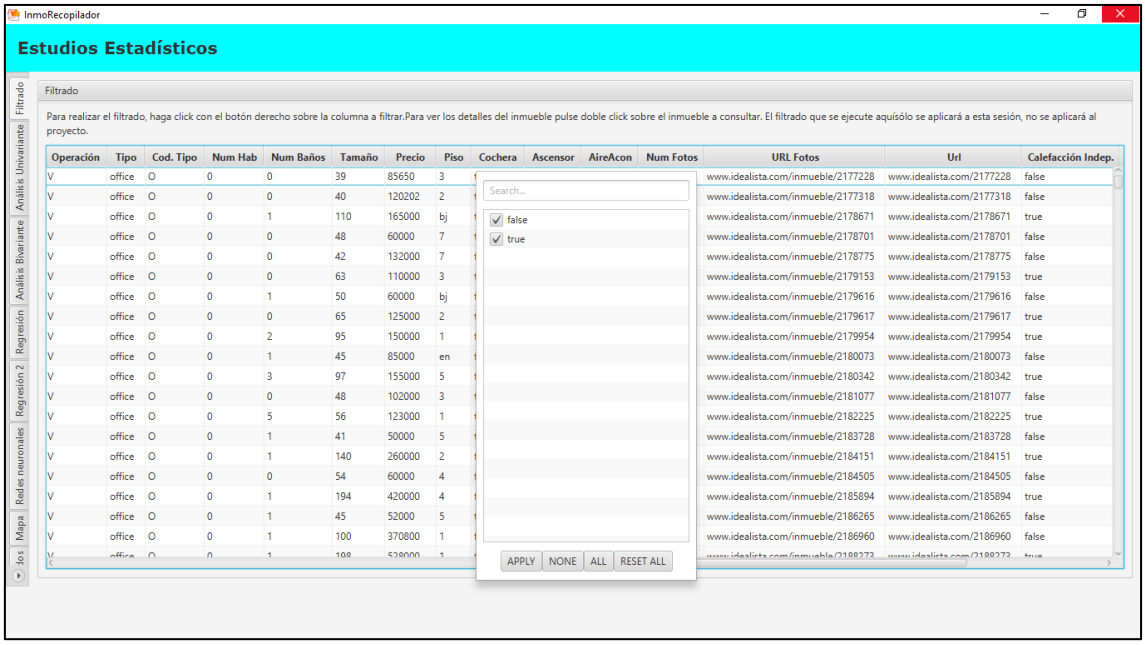


Figura 80. Pestaña de filtrado

6.6.5.2. La pestaña Análisis Univariante

Esta pestaña permite la realización de un estudio descriptivo, una o varias representaciones gráficas, el cálculo del intervalo de confianza para la media, así como estudios de normalidad de una o varias variables cuantitativas del proyecto activo.

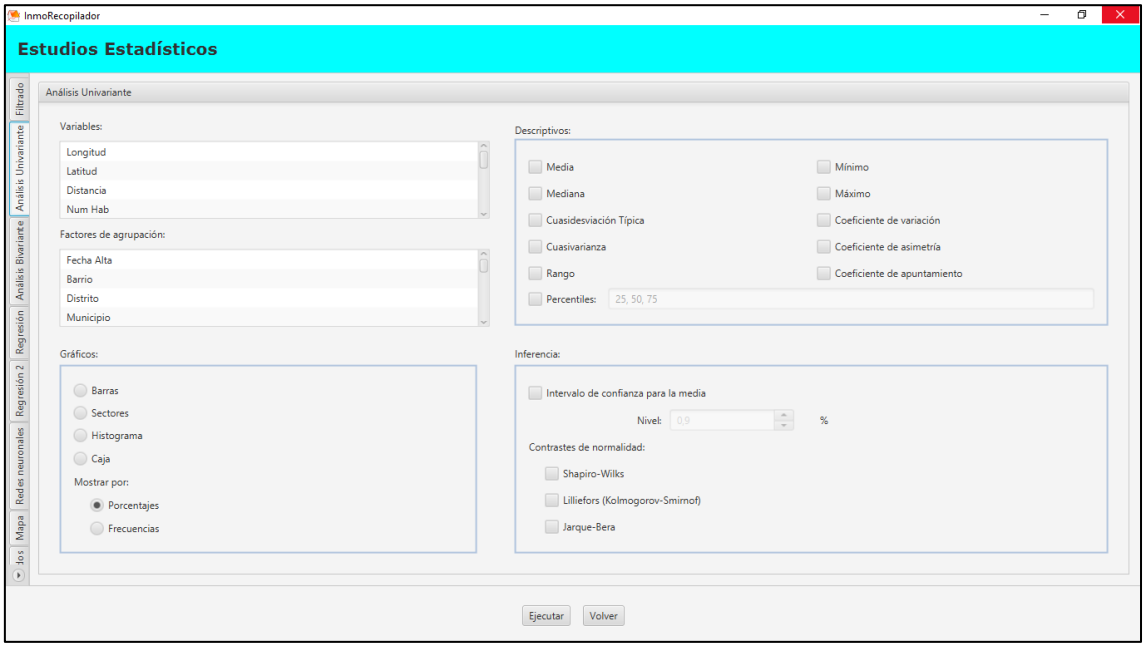


Figura 81. Pestaña de Análisis Univariante

Como puede observarse en la Figura 81, la ventana está dividida en cuatro bloques rectangulares.

El bloque superior izquierdo contiene, divididas en dos categorías, las variables para las que se puede realizar el estudio estadístico. La primera categoría, nombrada como la de las variables, contiene las variables cuantitativas o numéricas de la variable como son *Longitud*, *Latitud*, *Distancia*, *Tamaño* o *Precio*. La segunda contiene los denominados factores de agrupación, que son variables de tipo cualitativo, como son *Barrio*, *Distrito* o *Municipio*. El usuario tendrá que seleccionar obligatoriamente una o varias *variables*, y podrá seleccionar algún *factor de agrupación* o no.

En el caso de que se seleccionen únicamente variables de la primera categoría, el programa devolverá un estudio relativo a las variables elegidas. No obstante, si el usuario selecciona, además, variables de la segunda categoría, el programa devolverá un estudio en el que se divide cada variable cuantitativa en tantos grupos como modalidades tenga cada uno de los factores elegidos, y se realiza el análisis para cada grupo.

Por ejemplo, en un proyecto de locales, si se selecciona la variable Precio y el factor Localización, que incluye las modalidades Street, Mezzanine y Shoppingcenter, el programa le devolverá un estudio para cada uno de los tres grupos en los que se dividirán los locales, es decir, se realizará un análisis estadístico para el precio de los locales situados a pie de calle, otro el precio de los situados a pie de calle y un último estudio de los precios de los inmuebles situados en el interior de un centro comercial.

En el bloque superior derecho se pueden seleccionar un total de once estadísticos descriptivos que son:

- Medidas de posición: Media, mediana y percentiles. Al seleccionar esta última opción, se activará en el cuadro de texto en el que el usuario podrá escribir, separados por comas, qué percentiles desea que sean calculados, por defecto se incluyen los cuartiles.
- Medidas de dispersión: Valor mínimo, valor máximo, rango, cuasidesviación típica, cuasivarianza y coeficiente de variación.
- Medidas de forma: Coeficiente de asimetría y coeficiente de apuntamiento o curtosis, ambos de Fisher.

El bloque inferior derecho es el reservado al estudio inferencial de los datos. En él se puede calcular el intervalo de confianza para la media con un nivel de confianza que deberá seleccionar el usuario entre el valor mínimo que es 0.85 y el máximo, 0.99; a intervalos de una centésima.

También se puede realizar uno o varios contrastes que permitan determinar si los valores de la variable seleccionada son compatibles con que ésta provenga de una distribución Normal. El programa devolverá el valor del estadístico de contraste y su probabilidad límite asociada. Los contrastes disponibles son:

- Shapiro – Wilks.
- Lilliefors (Kolmogorov – Smirnov).
- Jarque – Bera.

Por último, en el bloque inferior izquierdo, el usuario dispone de las opciones necesarias para incluir en el estudio una representación gráfica. Éste podrá elegir entre:

- Diagrama de barras.
- Gráfico de sectores.
- Histograma.
- Diagrama de caja.

Además, también se podrá elegir si la escala de frecuencias de los gráficos se realiza a través de los porcentajes de las mismas o si por el contrario se desea utilizar las frecuencias absolutas de la variable.

Una vez seleccionadas todas las opciones deseadas, basta hacer clic en el botón *Ejecutar*, y se realizará el estudio solicitado. Los resultados se muestran en la pestaña *Resultados* de la ventana de *Estudios Estadísticos*. En la Figura 82, puede observarse un ejemplo de estudio univariante realizado para un conjunto de locales comerciales para los que se ha solicitado un análisis univariante del tamaño, el precio de venta y el número de fotografías del anuncio en la web. Los resultados se muestran redondeados a cuatro cifras decimales.

Análisis Univariante. Resumen.

Estudio descriptivo.

	Tamaño	Precio	NumFotos
Media	255,8308	264724,6654	5,6015
Mediana	132,0000	140000,0000	4,0000
Cuasidesviación típica	927,5529	463431,3923	6,0594
Cuasivarianza	860354,4128	214768655353,4914	36,7161
Mínimo	12,0000	9000,0000	0,0000
Máximo	13687,0000	4795400,0000	37,0000
Rango	13675,0000	4786400,0000	37,0000
Coeficiente de variación	3,6256	1,7506	1,0817
Coeficiente de asimetría	12,2152	5,9919	2,3322
Coeficiente de apuntamiento	166,0386	45,1803	6,8130
Ext. inf. IC media 95 %	143,8525	208777,1630	4,8700
Ext. sup. IC media 95 %	367,8091	320672,1679	6,3330
Normalidad (J-B) Estadístico	312169,3168	24215,6353	755,5957
Prob. límite	0,0000	0,0000	0,0000
Percentil 25	80,0000	78000,0000	2,0000
Percentil 50	132,0000	140000,0000	4,0000
Percentil 75	217,0000	269225,0000	7,0000

Figura 82. Ejemplo de resultado de análisis univariante. Tabla. (Fuente: Elaboración propia).

Como puede observarse, los distintos estudios se incluyen en la misma tabla, de forma que cada columna de ésta se corresponde con una de las variables seleccionadas.

La Figura 83 muestra el diagrama de caja de la variable NumFotos antes comentada. Este gráfico puede ser copiado en el portapapeles de Windows y posteriormente agregado a cualquier documento.

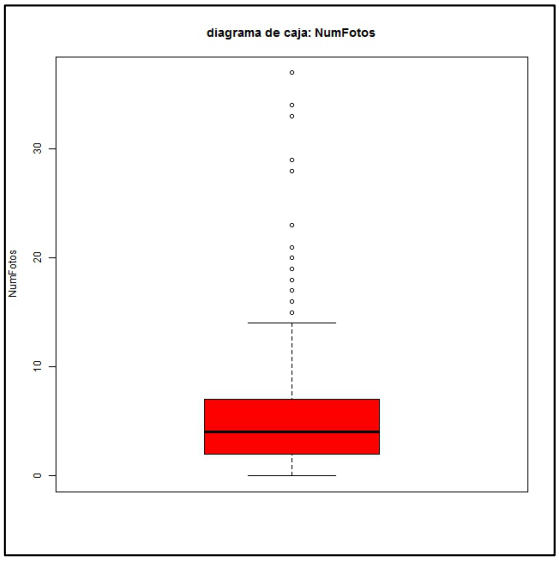


Figura 83. Ejemplo de resultado de análisis univariante. Gráfico. (Fuente: Elaboración propia).

La tabla de resultados difiere de la vista en la Figura 82 cuando se incluyen factores de agrupación en el estudio estadístico, ya que, en este caso, cada columna de la tabla se corresponde con un grupo diferente del factor. En la Figura 84 puede observarse la tabla resultante de realizar un análisis exploratorio del precio de un conjunto de viviendas en

función del estado del inmueble. Cada columna de ésta expresa los resultados para cada estado posible de la vivienda.

Análisis exploratorio.

Estudio descriptivo.

Precio x Estado

	good	newdevelopment	renew
Tamaño muestral	838,0000	16,0000	126,0000
Media	203545,9797	261815,6250	199512,5397
Mediana	162586,5000	261365,0000	150000,0000
Cuasidesviación típica	131640,0555	116195,7520	160832,4529
Cuasivarianza	17329104217,9339	13501452772,9167	25867077916,6984
Coeficiente de variación	0,6467	0,4438	0,8061
Coeficiente de aimetría	1,7862	0,0649	2,2873
Coeficiente de apuntamiento	3,5336	-1,8544	6,6170
Ext. inf. IC media 92 %	195575,1793	207269,3110	174222,9519
Ext. sup. IC media 92 %	211516,7801	316361,9390	224802,1275
Normalidad (S-W) Estadístico	0,8236	0,8474	0,7647
Prob. límite	0,0000	0,0125	0,0000
Percentil 17	100000,0000	150327,0000	80000,0000
Percentil 68	222794,1600	354000,0000	204000,0000

Figura 84. Ejemplo de resultado de análisis univariante. Tabla de análisis exploratorio. (Fuente: Elaboración propia).

Los gráficos difieren también de los anteriores, debido a que son representados, siempre que sea posible, en un mismo espacio todos los de los diferentes grupos realizados para una misma variable. Si el número de grupos en el que se divide la variable es superior a doce, se realizarán histogramas individuales debido a que, en ese caso, el área de trabajo disponible para cada gráfico es muy reducida.

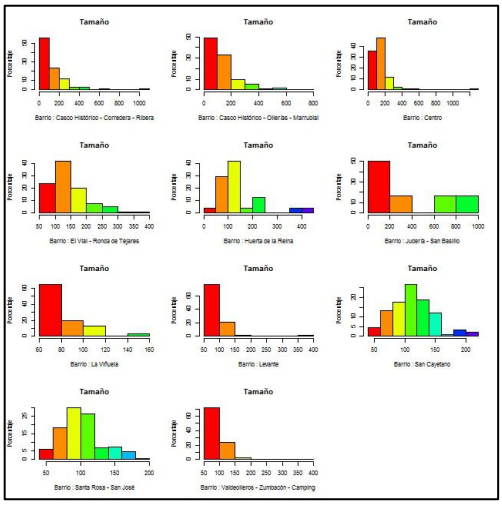


Figura 85. Ejemplo de resultado de análisis univariante. Histograma exploratorio. (Fuente: Elaboración propia)

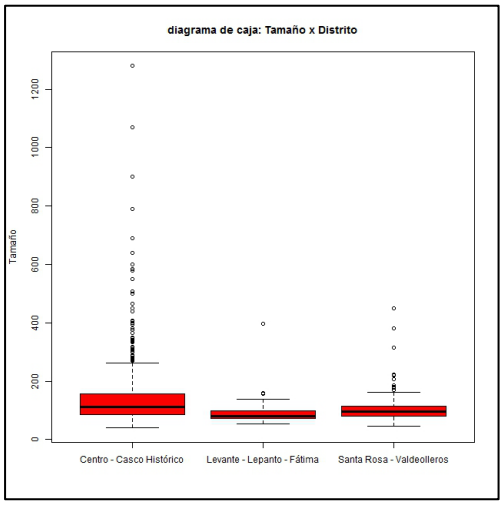


Figura 86. Ejemplo de resultado de análisis univariante. Diagrama de caja exploratorio. (Fuente: Elaboración propia)

En la Figura 85 se muestran los histogramas de la variable *Tamaño* de una vivienda para los 11 barrios de la ciudad de Córdoba analizados. En la Figura 86, para el mismo conjunto de datos se muestran los diagramas de caja de los tres distritos para los que se ha obtenido esta información.

6.6.5.3. La pestaña de Análisis bivariante

Esta pestaña permite la realización de dos estudios diferentes que tienen como objetivo analizar la relación existente entre dos variables. El menú correspondiente al primero de los estudios, el del análisis del grado de asociación entre variables cualitativas, se sitúa en la parte superior de la ventana, como puede observarse en la Figura 87. En la segunda mitad de la ventana se puede solicitar el estudio del grado de relación lineal entre variables cuantitativas.

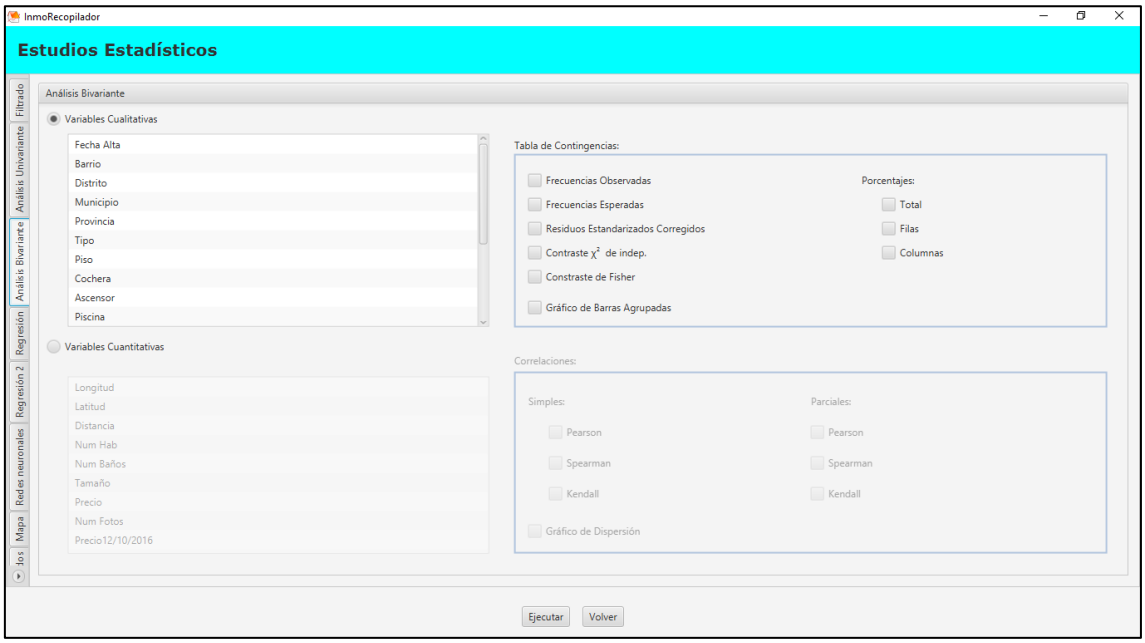


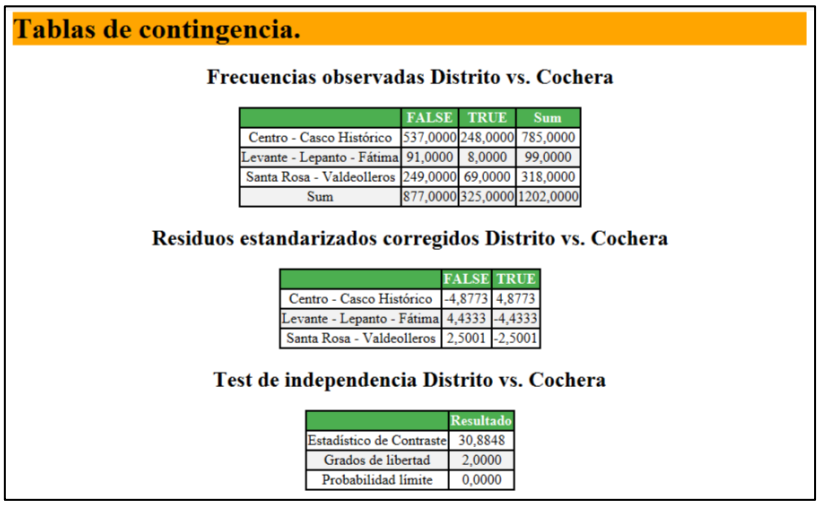
Figura 87. Pestaña de Análisis bivariante

En primer lugar, según el estudio que desee realizar, el usuario seleccionará una la opción *Variables Cualitativas* y *Variables Cuantitativas*. Sólo estará activa la parte de la ventana correspondiente a la del estudio seleccionado. En cualquiera de ellas, deberá elegir dos o más variables. Esto lo conseguirá dejando pulsada la tecla *Ctrl* del teclado mientras realiza la selección.

El cuadro superior derecho permite indicar las opciones necesarias para realizar el estudio del grado de asociación entre variables cualitativas. Las opciones disponibles permiten la construcción de diferentes tablas de contingencia, contrastes de independencia entre variables e incluso la representación gráfica en barras agrupadas de las frecuencias de los distintos cruces entre modalidades. Más detalladamente:

- Tabla de frecuencias observadas.
- Tabla de frecuencias esperadas bajo la hipótesis de independencia entre las variables.
- Tabla de porcentajes totales.
- Tabla de porcentajes por filas.
- Tabla de porcentajes por columnas.
- Residuos estandarizados corregidos, que permite analizar la causa de la asociación, en caso de que ésta exista.
- Contraste Chi – cuadrado de independencia, con el que se realiza un test de hipótesis para determinar si puede afirmarse la existencia de asociación entre las variables.
- Prueba exacta de Fisher. Útil Contraste de hipótesis para el estudio de asociación entre variables dicotómicas.
- Gráfico de barras de una de las variables, agrupada por las modalidades de la otra.

Si se seleccionan más de dos variables se realizará este estudio para todas las combinaciones de parejas de variables elegidas. En la Figura 88 se muestran algunos resultados del estudio del grado de asociación entre el distrito en el que se encuentra una vivienda y el hecho de poseer o no cochera.



Tablas de contingencia.

Frecuencias observadas Distrito vs. Cochera

	FALSE	TRUE	Sum
Centro - Casco Histórico	537,0000	248,0000	785,0000
Levante - Lepanto - Fátima	91,0000	8,0000	99,0000
Santa Rosa - Valdeolleros	249,0000	69,0000	318,0000
Sum	877,0000	325,0000	1202,0000

Residuos estandarizados corregidos Distrito vs. Cochera

	FALSE	TRUE
Centro - Casco Histórico	-4,8773	4,8773
Levante - Lepanto - Fátima	4,4333	-4,4333
Santa Rosa - Valdeolleros	2,5001	-2,5001

Test de independencia Distrito vs. Cochera

	Resultado
Estadistico de Contraste	30,8848
Grados de libertad	2,0000
Probabilidad límite	0,0000

Figura 88. Ejemplo de resultado de análisis bivariante. Tablas de contingencia. (Fuente: Elaboración propia).

En el estudio bivariante entre variables cuantitativas, el objetivo es determinar la existencia de relación lineal entre variables cuantitativas. Para ello disponemos de tres coeficientes con sus correspondientes contrastes de hipótesis: los coeficientes de Pearson, Spearman y Kendall. En concreto, las opciones de análisis de que se disponen son:

- Coeficientes de correlación simple y parcial de Pearson.
- Coeficientes de correlación simple y parcial de Spearman.
- Coeficientes de correlación simple y parcial de Kendall.
- Gráfico de dispersión. Al seleccionar esta opción, se mostrará cada uno de los gráficos de dispersión correspondientes a cada combinación de parejas seleccionadas y la matriz de gráficos de dispersión.

Los coeficientes de correlación parcial permiten el cálculo del coeficiente de correlación lineal entre dos variables, una vez eliminado el efecto de interacción provocado por el resto de variables de estudio.

En la Figura 89 puede observarse el resultado del coeficiente de correlación lineal simple de Pearson para las variables *NumBaños*, *Tamaño* y *Precio* de un conjunto de viviendas. La primera de las tablas muestra el valor de los coeficientes, la segunda muestra los estadísticos de los correspondientes contrastes de estudio de relación lineal de Pearson, y la tercera, las probabilidades límite asociadas a cada contraste.

Estudio de Correlación lineal.

Estimación del coeficiente de correlación de Pearson

	NumBaños	Tamaño	Precio
NumBaños	1,0000	0,6948	0,6900
Tamaño	0,6948	1,0000	0,7419
Precio	0,6900	0,7419	1,0000

Contraste de independencia de Pearson: Estadísticos

	NumBaños	Tamaño	Precio
NumBaños	∞	33,4917	33,0474
Tamaño	33,4917	∞	38,3583
Precio	33,0474	38,3583	∞

Contraste de independencia de Pearson: P - límite

	NumBaños	Tamaño	Precio
NumBaños	0,0000	0,0000	0,0000
Tamaño	0,0000	0,0000	0,0000
Precio	0,0000	0,0000	0,0000

Figura 89. Ejemplo de resultado de análisis bivariante. Correlación. (Fuente: Elaboración propia).

Como se ha comentado anteriormente, al seleccionar el gráfico de dispersión, se muestran como resultados todos los gráficos individuales, pero también la matriz de

gráficos de todas las variables. Ésta última y el gráfico de dispersión entre el Precio de la vivienda y su Tamaño se muestran en la *Figura 90* y *Figura 91*.

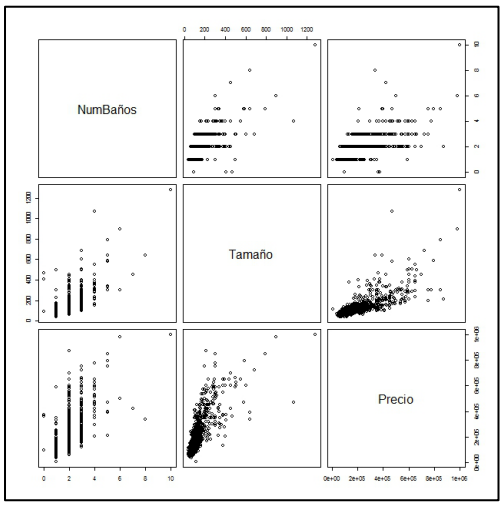


Figura 90. Ejemplo de resultado de análisis bivariante. Matriz de dispersión. (Fuente: Elaboración propia).

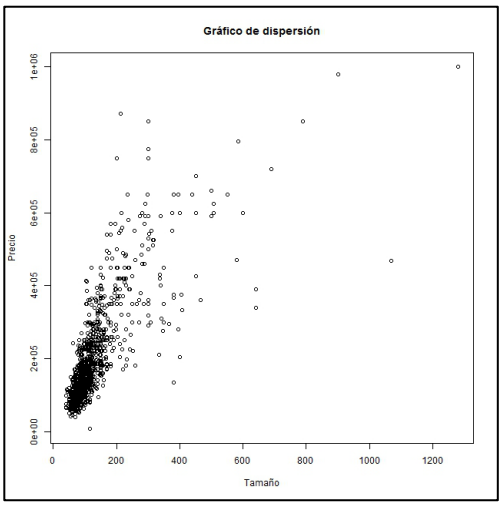


Figura 91. Ejemplo de resultado de análisis bivariante. Dispersión Tamaño-Precio. (Fuente: Elaboración propia).

6.6.5.4. La pestaña Regresión

La ventana de Regresión está dividida verticalmente en dos partes diferenciadas.

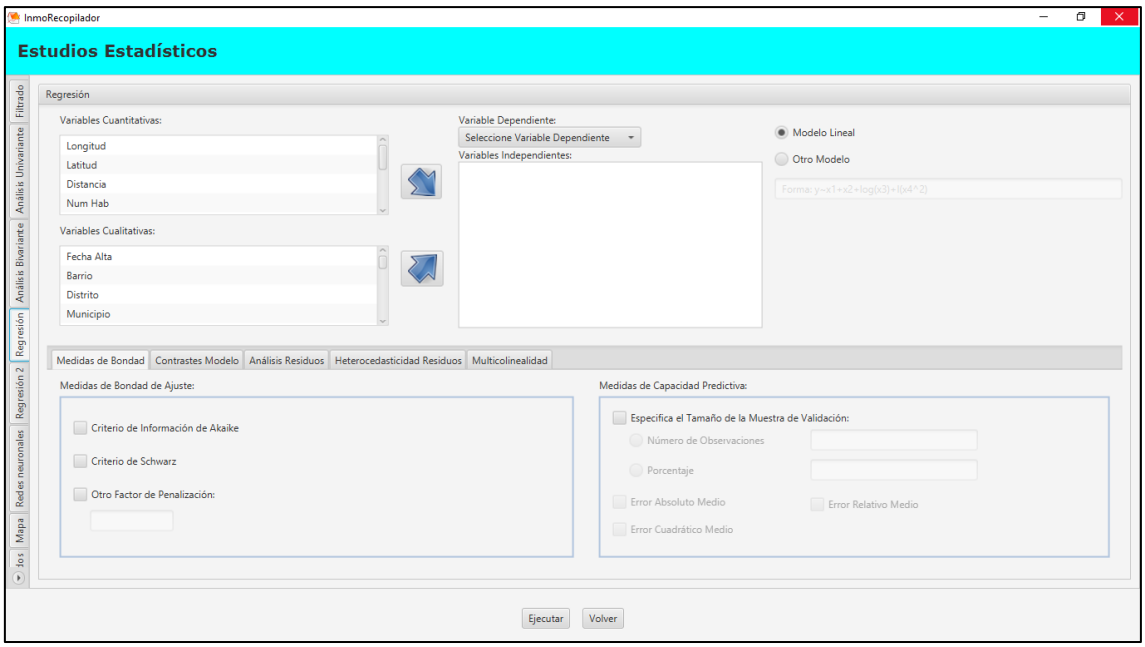


Figura 92. Pestaña de Regresión

La parte superior contiene los datos referentes a la construcción del modelo de regresión, y es fija. La parte inferior de la ventana se divide a su vez en cinco pestañas que contienen todas las funcionalidades de que dispone, como puede observarse en la Figura 92.

El modelo de regresión contiene una variable dependiente, que es la variable que se desea estimar, y que se seleccionará entre el grupo de variables cuantitativas asociadas al tipo de inmueble analizado. Esta variable es seleccionada a través del cuadro desplegable titulado Variable Dependiente, que se encuentra en la parte superior central de la ventana. Al hacer clic en la parte derecha se desplegará una lista con las variables dependientes disponibles, tal y como se puede observar en la Figura 93. Una vez seleccionada una variable de la lista, ésta no deberá ser seleccionada como variable independiente. Si esta acción es llevada a cabo, la selección realizada de la variable dependiente se eliminará, y al desplegar nuevamente el listado de opciones, la variable incluida como independiente no formará ya parte del listado.

Figura 93. Menú desplegable de selección de variable dependiente

En la parte superior izquierda de la ventana, divididas en dos grupos: *Variables Cuantitativas* y *Variables Cualitativas*, se listan todas las variables que pueden formar parte del modelo de regresión. Siempre deberá ser seleccionada al menos una variable cuantitativa del listado. Las variables cualitativas son incluidas en el modelo de regresión a través de variables indicadoras o *Dummies*. Para cada variable cualitativa, se introducirán en el modelo de regresión tantas variables *Dummies* como número de modalidades contenga la variable, es decir, una para cada modalidad; salvo para la última, que se excluye para evitar problemas de multicolinealidad exacta en el modelo.

Por ejemplo, para la variable Estado del grupo de viviendas, que puede tomar tres valores: *good*, *newdevelopment* y *renew*; se generarán dos variables *Dummies*, la variable *Estadogood*, que tomará el valor 1 para las viviendas que tomen el valor *good* en la variable *Estado* y 0 para el resto de viviendas, y la variable *Estadonewdevelopment* que será 0 salvo cuando la vivienda es de nueva construcción.

Para añadir variables al cuadro de variables independientes del modelo, basta con seleccionar una o varias (con la tecla Ctrl pulsada) de ellas, y hacer clic en el botón que contiene una flecha que señala dicho cuadro de texto.

Para el proceso contrario, es decir, para extraer una variable de entre el grupo de variables seleccionadas, basta con hacer clic sobre ella y pulsar el botón con la flecha al que se ha hecho referencia anteriormente. Éste cambiará la dirección de la flecha para indicar que se extraerá la variable seleccionada.

En la parte superior derecha, el usuario puede seleccionar la forma en la que intervienen las variables seleccionadas para formar el modelo de regresión. Por defecto, el modelo elegido es el lineal, esto es, un modelo de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

En el que $X_1, X_2, \ldots, X_k$ son las variables incluidas en el cuadro de *variables independientes*.

No obstante, es posible seleccionar modelos en los que se incluyan transformaciones de las variables seleccionadas, como el producto de dos o más de ellas o la potencia de una de las variables. Para esto, se deberá seleccionar la opción *Otro modelo*, y escribir éste con la siguiente formulación:

- La primera variable que aparece en la fórmula es la variable dependiente.
- A esta variable le sigue el símbolo ~, que puede interpretarse como el símbolo que indica que la variable a la izquierda de él es modelada a partir del grupo de variables a la derecha del mismo, unidas entre sí por el símbolo suma: +.
- El término $I(X\text{^}n)$ incluye en el modelo la potencia n-ésima de la variable X. También puede utilizarse, como base de la potencia, cualquier combinación de variables.
- El término $I(1/X)$ incluye en el modelo la inversa de la variable X.

- Dadas dos variables X_1 y X_2, el término $X_1 : X_2$ incluye la interacción entre ambas variables, es decir, la variable resultante de multiplicar ambas variables. Esto, puede escribirse equivalentemente como: $I(X_1 * X_2)$.
- En general, el término $I(\ )$ está indicado para distintas operaciones matemáticas entre variables.
- Es importante no confundir este último término con $X_1 * X_2$ que incluirá en el modelo cada una de las variables por separado y además su interacción.
- Dadas dos variables X_1 y X_2, si incluimos en el modelo el término $X_1|X_2$ se introducirá la variable X_1 condicionada a los valores de X_2.
- El término $poly(X, m)$ introduce, en el modelo, el polinomio ortogonal en X, de grado m.
- Si, además, se incluye en el lado derecho de la fórmula el término -1, se construirá el modelo de regresión sin término independiente.

Si se desea introducir, en la formulación del modelo, variables cualitativas, deberán incluirse con los nombres de las variables *Dummies* generadas, que se explicaron anteriormente. La introducción de una fórmula incorrecta generará un mensaje de error en el motor de R.

Si no se activa ninguna otra opción, salvo las correspondientes a la creación del modelo, los resultados que se mostrarán al hacer clic en el botón *Ejecutar*, estarán conformados por un total de tres tablas que resumen las principales características del modelo estimado. La Figura 94, muestra el resultado obtenido para un modelo de regresión que estima el Precio de un grupo de viviendas, mediante un modelo lineal, a partir del tamaño y la distancia al centro de la búsqueda.

Como puede observarse, la primera tabla muestra los valores del coeficiente de determinación y el ajustado, el estadístico del test de validación global del modelo, los grados de libertad asociados a éste y la probabilidad límite.

La segunda tabla muestra los valores estimados de los coeficientes, su error estándar, los estadísticos de contraste del test de relevancia de cada variable explicativa, y sus correspondientes probabilidades límite.

Por último, se muestra un resumen descriptivo de los residuos del modelo, que incluye el valor mínimo, el máximo, los cuartiles, la media y la cuasidesviación típica.

Estudio de Regresión.

Resultados de la estimación del modelo de regresión de la variable Precio

	Valores
Coeficiente de determinación	0,5506
Coeficiente de determinación ajustado	0,5498
Estadístico F	735,7052
Grados de libertad num.	2,0000
Grados de libertad den.	1201,0000
Probabilidad límite	0,0000

Coeficientes del modelo de regresión para la estimación de Precio

	Coeficientes	Error estándar	t	P-límite
Constante	62422,5753	8389,1467	7,4409	0,0000
Tamaño	1036,1930	27,0179	38,3520	0,0000
Distancia	8,0027	10,5838	0,7561	0,4497

Estadísticos descriptivos de los residuos

	Residuos
Mínimo	-710039,8898
Primer cuartil	-52904,7853
Mediana	-18981,7892
Media	0,0000
Tercer cuartil	32764,7103
Máximo	584639,0539
Desviación típica	91286,9263
Varianza	8333302913,0560

Figura 94. Ejemplo de resultado de regresión. (Fuente: Elaboración propia).

Una vez definido el modelo de regresión que se desea estimar, se dispone de un total de cinco pestañas, que dividen los análisis que pueden realizarse en diferentes categorías. Estas pestañas son:

- Medidas de bondad.
- Contrastes modelo.
- Análisis Residuos.
- Heterocedasticidad Residuos.
- Multicolinealidad.

6.6.5.4.1. Pestaña de Regresión: Medidas de Bondad

La pestaña *Medidas de bondad* permite el cálculo de dos tipos de medidas de la bondad del modelo construido en la parte superior. En la Figura 95 se puede observar que la pestaña está dividida en dos bloques:

- En el bloque izquierdo se presentan las medidas necesarias para determinar la capacidad del modelo para ajustarse a los datos. Estas son los criterios de información de Akaike y Schwarz, que se diferencian en el factor de penalización sobre el número de variables explicativas incluidas en el modelo. También es posible calcular un criterio de información con un factor de penalización definido por el usuario. Otras medidas de bondad de ajuste como el coeficiente de determinación son calculadas por defecto en la tabla de resultados de la estimación, como se ha mencionado anteriormente.

- Por otro lado, si se divide la muestra disponible en dos submuestras, de forma que la primera de ellas se utilice para estimar el modelo de regresión, la segunda se podrá utilizar para estimar la capacidad predictiva del modelo de regresión construido. Ésta última muestra se denomina muestra de validación. Por tanto, para el cálculo de las *Medidas de Capacidad Predictiva* es necesario, en primer lugar, indicar el tamaño de la muestra de validación. Para ello, se introducirá el número de observaciones o el porcentaje del total de éstas que se desean reservar para la validación del modelo. Una vez completado esto, se podrá calcular el error absoluto medio, el error relativo medio y el error cuadrático medio para la muestra de validación. Cabe destacar que los inmuebles que formarán parte de la muestra de validación serán seleccionados de forma aleatoria por el programa hasta completar el tamaño solicitado.

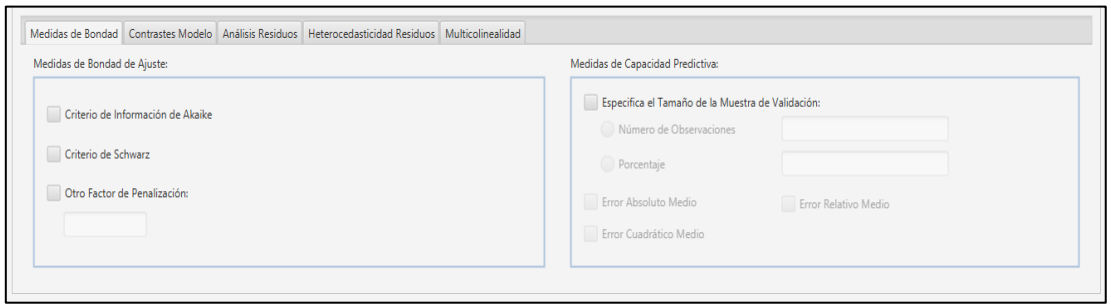


Figura 95. Regresión: Medidas de bondad

La Figura 96 muestra el resultado obtenido para las medidas de bondad de un modelo de regresión que se ha construido para estimar el precio de venta de un local en la ciudad de Córdoba a partir de su tamaño, el número de baños de que dispone y si está situado haciendo esquina entre dos calles. Como puede observarse los criterios de información se muestran en una primera tabla, indicando en el tercero el factor de penalización utilizado. En una segunda tabla se muestran los errores medios para la muestra de validación del tamaño seleccionado, que en este caso es de 64, que se corresponde con el 18 por ciento del total de la muestra.

Criterios de información

	Valores
C. I. de Akaike	9844,0240
C. I. Bayesiano (Schwarz)	9863,3705
C. I. con factor de penalización 3	9849,0240

Capacidad predictiva del modelo a partir de una muestra de validación de tamaño: 64

	Valores
Error absoluto medio	138208,8808
Error relativo medio	0,8781
Error cuadrático medio	133884784613,7103

Figura 96. Ejemplo de regresión. Medidas de bondad. (Fuente: Elaboración propia).

6.6.5.4.2. Pestaña de Regresión: Contrastes Modelo

La pestaña Contrastes Modelo permite contrastar, por un lado, la idoneidad de la forma funcional seleccionada, y por otro lado calcular intervalos de confianza para los coeficientes del modelo de regresión. Así, como se observa en la Figura 97, la pestaña se divide en dos bloques.

En el primero de ellos se analiza la linealidad del modelo seleccionado de dos formas: mediante el gráfico de residuos parciales y a través de la aplicación del contraste RESET. También se puede aplicar el contraste de Davidson – Mackinnon para comparar el modelo elegido con otro modelo con la misma variable dependiente, con diferente forma funcional, siempre que el conjunto de variables explicativas usadas, esté contenido en el de variables seleccionadas.

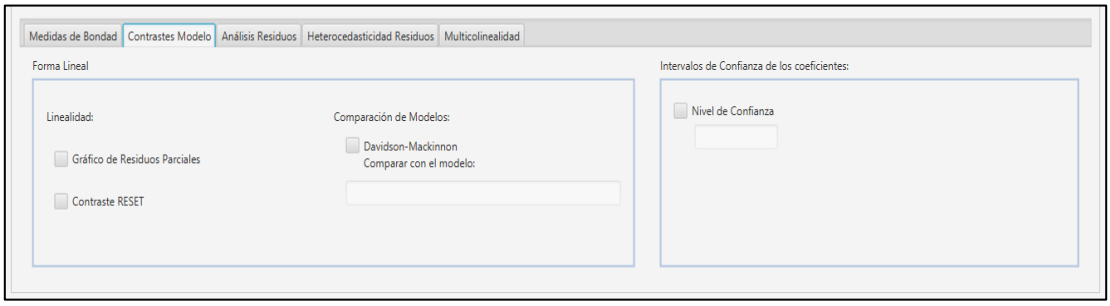


Figura 97. Regresión: Contrastes Modelo (Fuente: Elaboración propia)

El bloque de la derecha de esta pestaña sirve únicamente para la estimación, por intervalo de confianza, de los coeficientes del modelo de regresión. El nivel de confianza debe ser aportado por el usuario al programa. Los valores admitidos son los comprendidos entre 0.01 y 0.99, aunque no son recomendables valores por debajo de 0.90.

Para el caso anteriormente descrito del modelo para estimar los precios de los locales de la ciudad de Córdoba, se muestran, en la Figura 98, los intervalos de confianza al 95 por ciento para los coeficientes del modelo, así como el contraste *RESET* y el de *Davidson – Mackinnon* para la formulación *Precio~Tamaño*NumBaños*, es decir, se compara en éste último, el modelo que incluye *Tamaño*, *NumBaños* y *Esquina* con el que explica *Precio* a partir de *Tamaño*, *NumBaños* y su interacción.

El programa

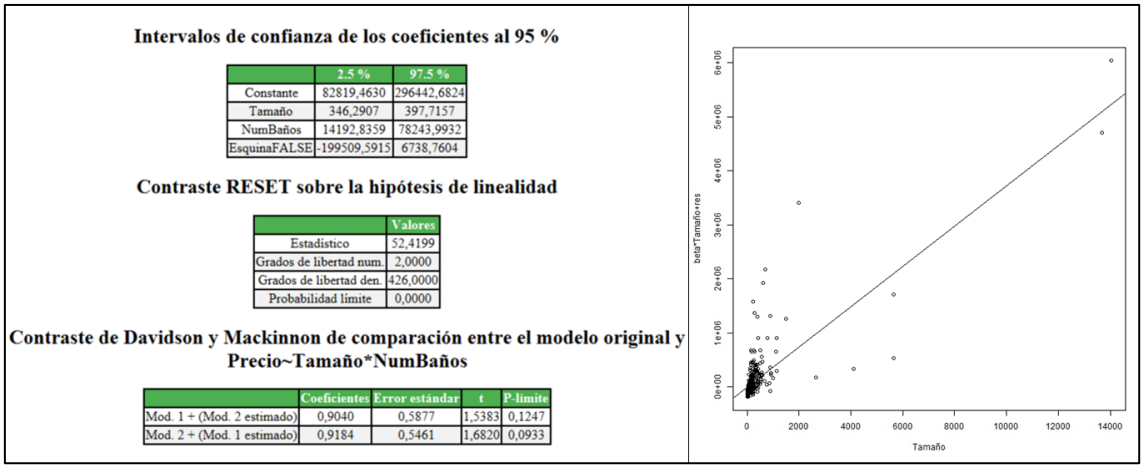


Figura 98. Ejemplo de regresión. Contrastes Modelo. (Fuente: Elaboración propia).

6.6.5.4.3. Pestaña de Regresión: Análisis residuos

Veamos a continuación la pestaña Análisis Residuos (Figura 99). En ella, se analiza el cumplimiento de dos hipótesis sobre los errores del modelo que son necesarias para la validación de éste. La normalidad y la ausencia de autocorrelación.

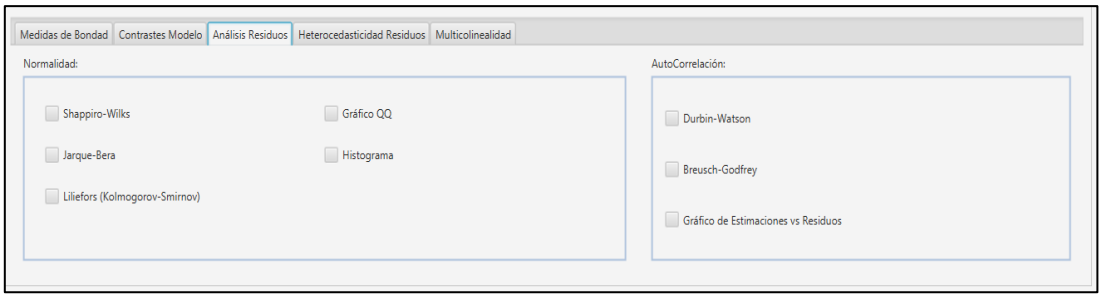


Figura 99. Regresión: Análisis Residuos

Para el análisis de la normalidad pueden utilizarse los tres contrastes de hipótesis vistos en la pestaña de Análisis Univariante: *Shapiro – Wilks*, *Jarque – Bera* y *Lillieffors*. También es posible representar un histograma y un gráfico QQ de los residuos. Ambos incluyen la línea teórica de normalidad para su mejor comparación con los resultados empíricos.

El estudio de la autocorrelación de los residuos se puede realizar gráficamente, a través del gráfico de dispersión entre los valores estimados por el modelo y los residuos de éste. También, por medio de dos contrastes de hipótesis, el de *Durbin – Watson* para

el estudio de autocorrelación de primer orden, y el de *Breusch – Godfrey* para la detección de autocorrelación de orden superior.

Los resultados de esta pestaña se muestran en dos tablas. La primera de ellas presenta los estadísticos de contraste y las probabilidades límite de las pruebas de normalidad seleccionadas. La segunda tabla muestra estos resultados para los contrastes de autocorrelación. Además, se realizará una representación por cada gráfico solicitado.

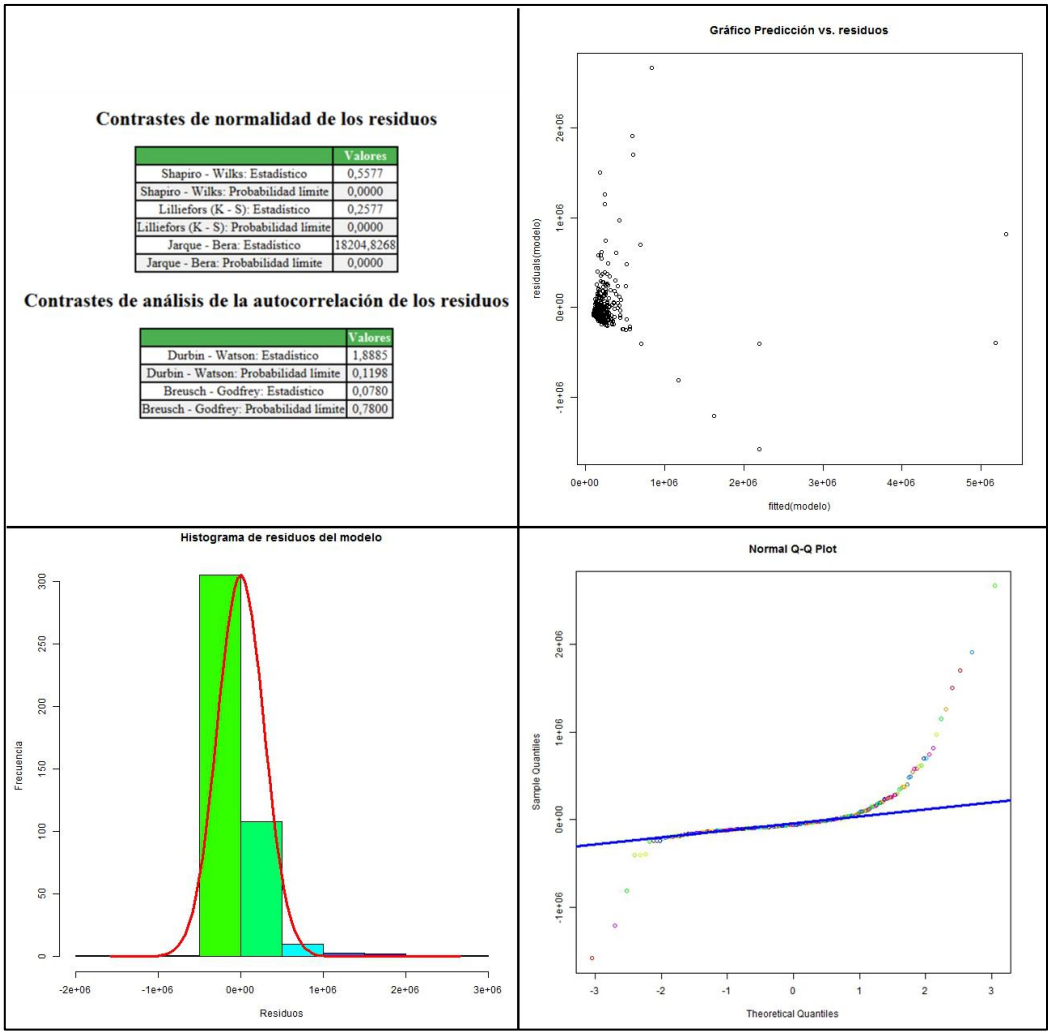


Figura 100. Ejemplo de regresión. Análisis residuos. (Fuente: Elaboración propia).

Retomando de nuevo el ejemplo del modelo de regresión para la estimación del precio de venta de los locales comerciales de la ciudad de Córdoba, sus resultados para las pruebas de hipótesis y los gráficos de los residuos contenidos en esta pestaña se pueden observar en la Figura 100, que se ha dividido en cuatro bloques, el primero de los cuales

corresponde a las tablas de resultados. Los otros tres se corresponden con las representaciones gráficas correspondientes.

6.6.5.4.4. Pestaña de Regresión: Heterocedasticidad

El análisis de los residuos del modelo de regresión se completa con el estudio de su heterocedasticidad. Esto puede hacerse desde la siguiente pestaña disponible de la opción *regresión*: *Heterocedasticidad Residuos*. Una imagen de ésta puede observarse en la Figura 101.

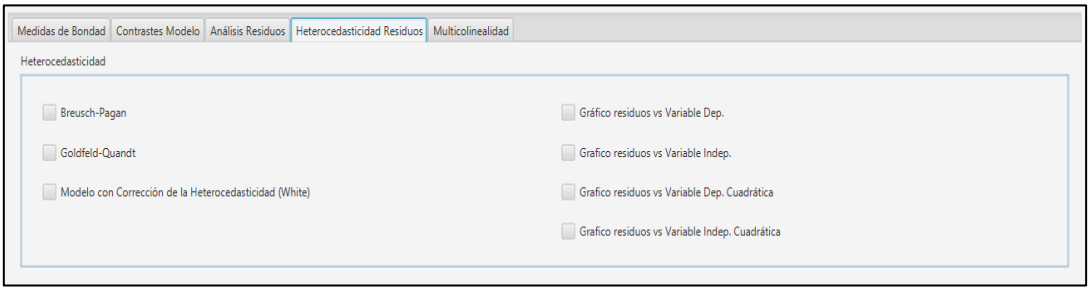


Figura 101. Regresión: Heterocedasticidad Residuos

Esta pestaña permite al usuario realizar dos contrastes para la determinación de ausencia de heterocedasticidad residual, el de Breusch – Pagan y el de Goldfeld – Quandt. El resultado de éstos se muestra en una tabla en la que se muestra el valor de los estadísticos de contraste, y sus correspondientes probabilidades límite.

Se ha omitido de estos contrastes el test de White, que históricamente es el más utilizado en el análisis de regresión, debido a la imposibilidad de su aplicación cuando se introducen en el modelo variables *dummies*, ya que el modelo auxiliar utilizado en este contraste incluye distintas potencias de las variables explicativas como variables de éste, lo que provoca multicolinealidad exacta entre estas, por ser las distintas potencias de una variable *dummy*, iguales a la variable original.

También permite cuatro tipos de representaciones gráficas que tienen como objetivo el análisis de las causas que de la heterocedasticidad. Éstas son:

- *Gráfico de residuos vs. Variable Dep*. Es un gráfico de dispersión en el que se sitúan en el eje de abscisas los valores de la variable dependiente, y en el de ordenadas los residuos del modelo.

- *Gráfico de residuos vs. Variable Indep*. Se representan *k* gráficos de dispersión, uno por cada variable explicativa que contenga el modelo. En éstos, se sitúan, de nuevo, en el eje de ordenadas los residuos del modelo, y en el de abscisas, los valores de cada una de las variables independientes. A partir de estos gráficos puede determinarse la variable regresora que, en su caso, provoca la existencia de heterocedasticidad.
- *Gráfico de residuos cuadráticos vs. Variable Dep*. Es similar al primero de los gráficos comentados, con la diferencia de que en el eje de ordenadas se representan los cuadrados de los errores, que es una forma de observar la dispersión de los residuos del modelo.
- *Gráfico de residuos cuadráticos vs. Variable Indep*. De forma equivalente a los anteriores, se definen éstos, representando los errores al cuadrado en el eje de ordenadas y cada variable independiente en el de abscisas.

Por último, se ofrece la posibilidad de construir, a través del método de mínimos cuadrados generalizado, la estimación del modelo, con corrección de heterocedasticidad; utilizando como estimación de las varianzas residuales, las dadas por el método de White. Como se ha comentado anteriormente, la introducción de variables dummies en el modelo, imposibilita el cálculo de esta estimación. Esto es indicado en el resultado devuelto.

En caso de ser posible la estimación por el método de mínimos cuadrados generalizado, ésta es devuelta a través de dos tablas, en la primera de ellas se muestran los coeficientes del modelo y sus errores estándares corregidos; así como el contraste de relevancia de éstos. También se vuelven a calcular los intervalos de confianza de los coeficientes.

Contrastes de análisis de la heterocedastiidad de los residuos

	Valores
Breusch - Pagan Estudentizado: Estadístico	25,3833
Breusch - Pagan Estudentizado: Probabilidad límite	0,0000
Goldfeld - Quandt: Estadístico	0,7833
Goldfeld - Quandt: Probabilidad límite	0,9541

Modelo con corrección de la heterocedasticidad (White)

	Coeficientes	Error estándar	t	P-límite
Constante	95480,3619	20409,2069	4,6783	0,0000
NumBaños	54014,6517	26404,2768	2,0457	0,0414
Tamaño	371,7992	50,1590	7,4124	0,0000

Intervalos de confianza de los coeficientes con la corrección de la heterocedasticidad

	50 %	50 %
(Intercept)	95480,3619	95480,3619
NumBaños	54014,6517	54014,6517
Tamaño	371,7992	371,7992

Figura 102. Ejemplo de regresión. Heterocedasticidad residuos. (Fuente: Elaboración propia).

Las tablas obtenidas en el análisis de heterocedasticidad para el ejemplo explicado anteriormente, al que se ha eliminado la variable Esquina para poder mostrar el modelo con corrección de heterocedasticidad, pueden observarse en la Figura 102.

6.6.5.4.5. Pestaña de Regresión: Multicolinealidad

La última opción de la pestaña de regresión, se reserva para el estudio de la multicolinealidad entre las variables explicativas del modelo de regresión. Como se puede ver en la Figura 103, se da la opción al usuario de medir el grado de multicolinealidad existente a través del cálculo de los factores de inflación de la varianza o *FIV* o a través del número de condición de la matriz de datos, también para la matriz de datos normalizada.

También es posible, al igual que en la pestaña anterior, construir un modelo que incluya la corrección de multicolinealidad. En este caso, el modelo seleccionado es el de Ridge, que incluye un término constante en su formulación que evita la singularidad de la matriz de construcción del modelo. El valor de esta constante será aportado por el usuario, y podrá tomar cualquier valor comprendido entre 0 y 1.

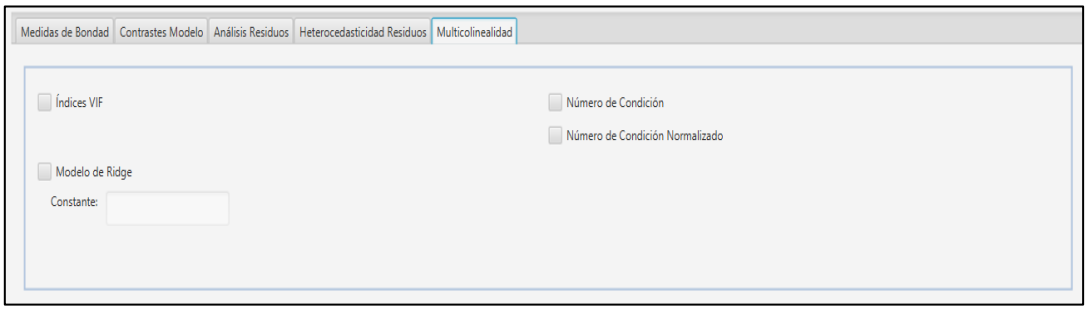


Figura 103. Regresión: Multicolinealidad

Los valores de los FIV, para cada variable independiente, se muestran en una tabla, el número de condición y el normalizado se muestran en otra, y los coeficientes estimados para el modelo de regresión por el método de White se muestran en una tercera tabla.

Pueden verse en la Figura 104, estos resultados para el ejemplo de los locales comerciales de la ciudad de Córdoba, en el que han vuelto a incluirse la variable Esquina, que se eliminó en el estudio de la heterocedasticidad para poder incluir el modelo corregido de White.

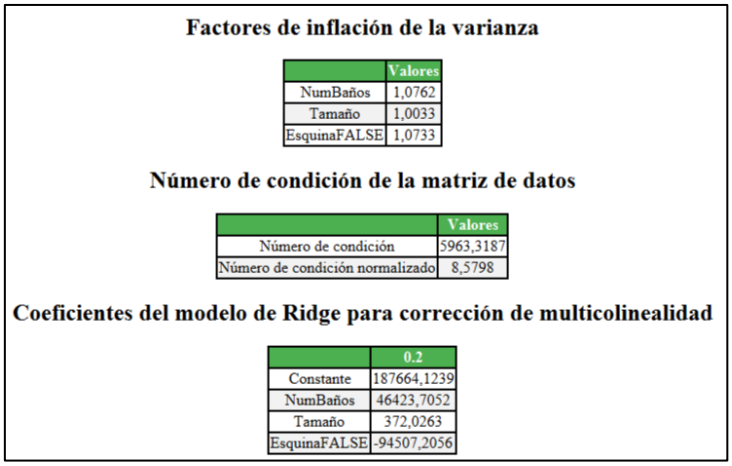


Figura 104. Ejemplo de regresión. Multicolinealidad. (Fuente: Elaboración propia).

6.6.5.5. La pestaña Regresión 2

La estructura de la ventana *Regresión 2* es similar a la de la ventana *Regresión*. Está dividida en dos bloques horizontalmente. La parte superior, en la que se realiza la formulación del modelo de regresión, que es igual a la de la ventana anterior, y la parte inferior de la ventana, subdividida en pestañas, con las diferentes opciones a seleccionar.

El bloque superior es idéntico al utilizado para la formulación del modelo en la ventana *Regresión*, y fue descrito en la sección anterior.

El objetivo de esta ventana es la de identificar la existencia de inmuebles con características diferenciadoras, en algún sentido, del resto. Existen multitud de métodos que pueden utilizarse para la identificación de observaciones atípicas o también denominadas *outliers*. En esta ventana se pone a disposición del usuario algunos de ellos, e incluso se propone un nuevo método para la detección de este tipo de observaciones.

Como se puede observar en la Figura 105, la parte inferior de la ventana está subdividida en cuatro secciones, a las que se accede a través de igual número de pestañas.

El programa

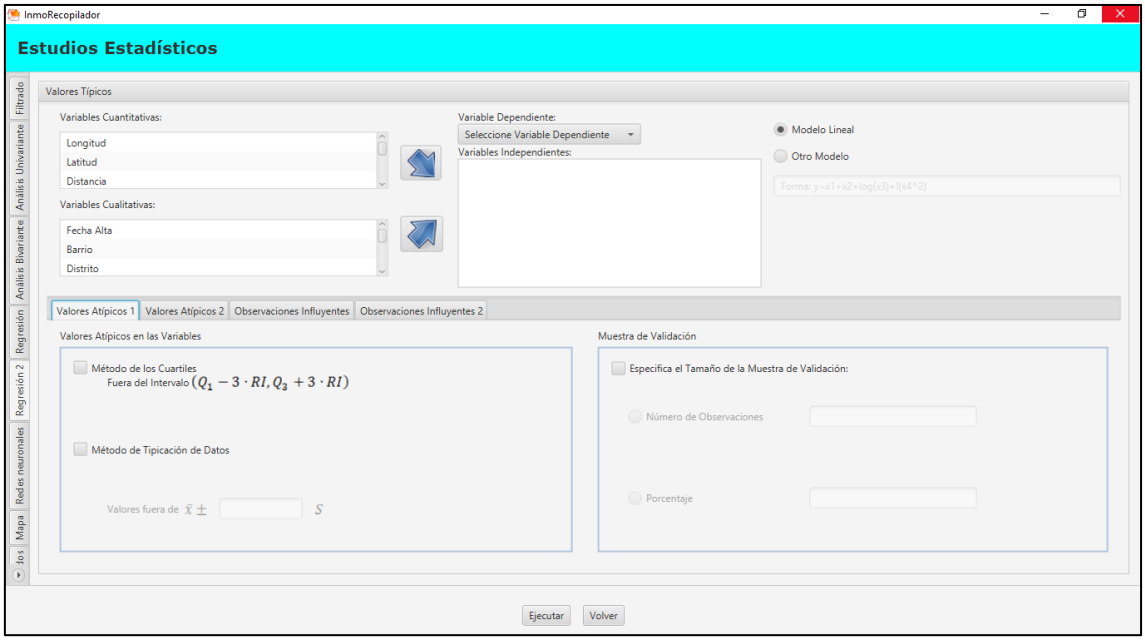


Figura 105. Pestaña de Regresión 2

Éstas son:

- Valores atípicos 1.
- Valores atípicos 2.
- Observaciones influyentes 1.
- Observaciones influyentes 2.

Como se ha comentado en otro capítulo, los métodos de detección de valores atípicos pueden dividirse en dos categorías: los métodos de tipo univariante, que para cada observación analizan el valor de todas las variables, identificando valores muy alejados de los obtenidos para el resto de observaciones en alguna de las variables estudiadas; y los métodos multivariantes, que analizan la observación en su conjunto y su influencia sobre la bondad de ajuste del modelo de regresión estimado.

6.6.5.5.1. Pestaña de Regresión 2: Valores atípicos 1.

En la pestaña Valores Atípicos 1, se divide a su vez en dos bloques verticalmente. El bloque derecho, con el que el usuario puede dividir la muestra en dos: la muestra de estudio y la de validación, y la de la izquierda, con la que puede determinarse la existencia de valores atípicos en las variables incluidas en el modelo de la parte superior del cuadro de diálogo.

El bloque derecho, al igual que en la pestaña Regresión, se utilizará para indicar el porcentaje o el número de inmuebles que desean dejarse para la muestra de validación. Estas observaciones son elegidas aleatoriamente de entre la muestra de estudio.

En el bloque izquierdo, el usuario podrá seleccionar el método con el que se identificarán las observaciones atípicas y el criterio a seguir.

El método de los cuartiles identificará, para todas las variables cuantitativas del modelo, tanto la variable dependiente como las independientes, qué inmuebles pueden considerarse outliers. Identificará como tales, los valores que se desvíen de los cuartiles uno y tres, por defecto y exceso respectivamente, tres veces el rango intercuartílico.

Por otro lado, el método de tipificación de datos identifica, en cada variable de forma independiente también, los valores atípicos, como aquellos que se desvían de la media aritmética un número de veces, determinado por el usuario, la desviación típica de la variable.

Se ha seleccionado, como ejemplo, un conjunto de locales comerciales a la venta de la ciudad de Jaén. La variable dependiente ha sido el precio de venta del inmueble, y las variables independientes: el tamaño, el número de baños y el número de fotografías del anuncio.

El resultado obtenido para la detección de outliers, en este ejemplo con un factor de desviación de tres, es el que se muestra en la Figura 106. En la imagen de la derecha se muestran los valores atípicos por el método de los cuartiles, y en la de la izquierda, los detectados por el método de tipificación de datos. Como puede observarse, para cada variable se indican las observaciones que han sido consideradas anómalas, identificándose su valor y el código de inmueble en la web del proveedor de los datos, que como se indicó con anterioridad es único para cada inmueble.

En caso de que no se hallen valores atípicos, un texto indica este resultado, como puede observarse para el número de baños del local comercial en el método de los cuartiles.

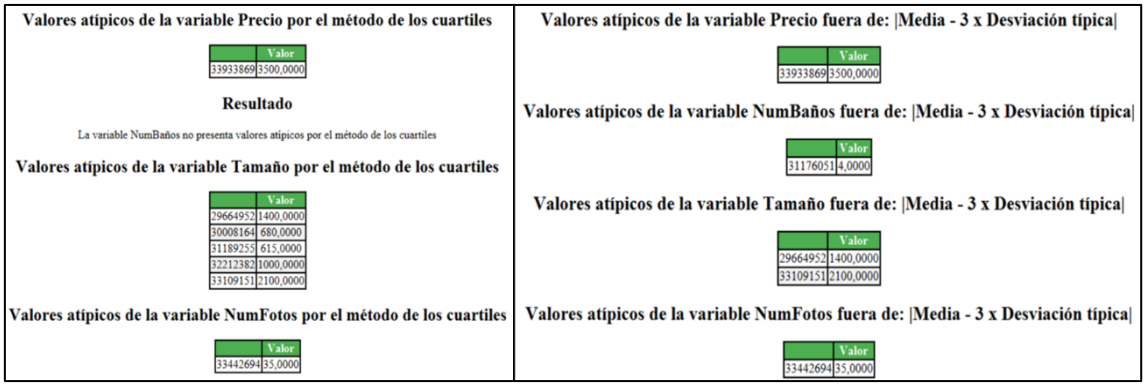


Figura 106. Ejemplo de regresión 2. Valores atípicos 1. (Fuente: Elaboración propia).

6.6.5.5.2. Pestaña de Regresión 2: Valores atípicos 2.

La pestaña *Valores Atípicos 2* permite analizar la presencia de outliers mediante el estudio de los residuos del modelo de regresión definido, y permite realizar representaciones gráficas para analizar las desviaciones de éstos y medir la influencia de cada observación en la estimación del modelo.

En primer lugar, y usando de nuevo el método de tipificación, se analizan los residuos del modelo que pueden considerarse atípicos. La observación que da lugar a un residuo anómalo es mal estimada por el modelo de regresión y por tanto puede considerarse anómala. Es también, en esta ocasión, el usuario el que decide la constante de desviación utilizada. Puede ser utilizado cualquier valor positivo.

El segundo método para la detección de valores atípicos es el test de Bonferroni, que decide si un inmueble es considerado como tal a partir del estudio de los residuos estudentizados. El usuario podrá introducir el nivel de significación con el que desea realizar el contraste. Valores de probabilidad límite por debajo del propuesto se corresponderán con observaciones atípicas. El valor propuesto por defecto es 0.05.

En la parte derecha de la pestaña, como puede observarse en la Figura 107, se encuentra el listado de gráficos disponibles, en el que el usuario podrá seleccionar los que se representarán. Estos gráficos son:

- Gráfico QQ de residuos estudentizados. La desviación respecto a los cuantiles teóricos arroja información sobre la existencia de observaciones anómalas.

- Gráfico de valores con efecto palanca. Se divide el área de representación y se representan tantos gráficos de dispersión como variables explicativas contenga el modelo. En todos ellos se sitúa en el eje de ordenadas la variable dependiente.
- Gráfico de Distancias de Cook. Gráfico de barras en el que el eje de abscisas se sitúan todas las observaciones, y a partir de cada una de ellas se levanta una barra con una altura proporcional a la distancia de Cook para dicha observación.
- Gráfico conjunto de medidas de influencia. Esta opción devuelve tres gráficos de barras similares al anterior, en los que las alturas de las barras representan, respectivamente, los residuos estandarizados, las probabilidades límite del test de Bonferroni, y los valores palanca de cada observación.

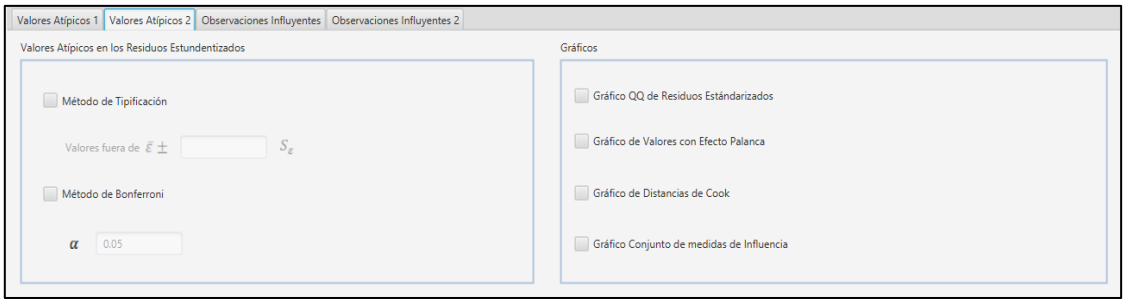


Figura 107. Regresión 2: Valores atípicos 2

Los residuos estudentizados considerados atípicos son identificados en el resultado mediante el código de su inmueble, y son situados en una tabla en la que además se incluye el valor de las variables independientes del modelo para una más rápida identificación del inmueble.

El resultado para el método de Bonferroni se muestra en una tabla en la que se identifica el inmueble por su código y se incluye además el valor del residuo estudentizado, la probabilidad límite no ajustada y la probabilidad límite de Bonferroni. En la tabla se incluye los resultados de los inmuebles para los que la probabilidad límite es inferior a la indicada por el usuario.

En la Figura 108, se muestra el resultado obtenido en la identificación de valores atípicos a partir de los residuos estudentizados del modelo para un conjunto de oficinas a la venta de la ciudad de Sevilla. Se muestra a la derecha, además, uno de los gráficos obtenidos, el gráfico de barras de los residuos estudentizados.

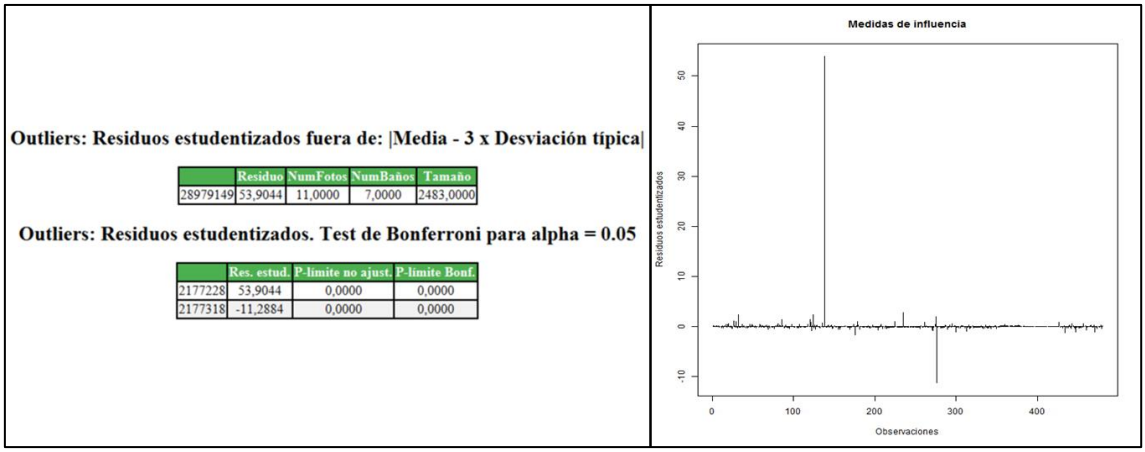


Figura 108. Ejemplo de regresión 2. Valores atípicos 2. (Fuente: Elaboración propia).

6.6.5.5.3. Pestaña de Regresión 2: Observaciones influyentes

La tercera pestaña es la de *Observaciones Influyentes*, Figura 109. A través de ella, el usuario puede analizar la presencia de observaciones influyentes mediante cinco métodos. En todos ellos se mostrarán los inmuebles que superen el límite permitido para cada uno de ellos. Se proponen los valores más utilizados en la literatura para estos límites. No obstante, podrá ser el usuario el que defina un valor distinto.

Los métodos de detección de observaciones influyentes disponibles son:

- Observaciones con efecto palanca. En este método se mide la distancia de una observación al resto de observaciones de la muestra. La identifica como influyente si la distancia supera el límite permitido.
- Dfbetas. En este método se mide la influencia que tiene cada observación sobre los coeficientes del modelo de regresión. Una observación es considerada influyente si al menos uno de los coeficientes del modelo sufre una desviación superior al límite permitido.
- Dffits. Mide la diferencia en el valor estimado por el modelo para una observación incluida en él y la obtenida si esta observación es excluida.
- Distancia de Cook. Similar al anterior, pero teniendo en cuenta todas las estimaciones, no sólo la de la observación estudiada.
- COVRatio. Mide el ratio entre el determinante de la matriz de covarianzas de los coeficientes del modelo obtenido al eliminar la observación analizada, y el obtenido conteniéndola.

Figura 109. Regresión 2: Observaciones influyentes 1

Para cada método seleccionado se obtendrá una tabla, en el caso de obtenerse observaciones influyentes, que contiene el código de cada uno de los inmuebles seleccionados, los estadísticos del método, el valor de la variable dependiente, y el de las independientes. Además, el título que precede cada tabla informará del valor o valores límite utilizados.

Se ha considerado el mismo ejemplo de oficinas a la venta en la ciudad de Sevilla que ya se comentó en la pestaña *Valores Atípicos 2*. Los resultados para diferentes valores límite especificados en el título que precede cada tabla se muestran en la Figura 110.

Medidas de influencia: Valores con un efecto palanca mayor que 0.1

	Efecto palanca	Precio	NumFotos	NumBaños	Tamaño
29965257	0,2387	4100000,0000	1,0000	0,0000	4637,0000
31928760	0,6014	4900000,0000	5,0000	4,0000	7500,0000
33720107	0,1119	44000,0000	12,0000	12,0000	60,0000

Medidas de influencia: Observaciones con valores DFFITS mayores que 0.2

	DFFITS	Precio	NumFotos	NumBaños	Tamaño
28979149	16,1020	12500000,0000	11,0000	7,0000	2483,0000
31408255	0,2385	2105000,0000	2,0000	1,0000	663,0000
31890683	0,2106	2105000,0000	8,0000	1,0000	1030,0000

Medidas de influencia: Observaciones con valores Dfbetas estandarizados mayores que 0.5

	Constante	NumFotos	NumBaños	Tamaño
28979149	-4,7488	-0,4415	7,8913	12,5379
29965257	-0,0341	0,1025	0,1955	-0,9406
31928760	1,5111	0,8245	0,3846	-13,7505

Medidas de influencia: Observacioes cuya distancia de Cook es mayor que 0.5

	D. Cook	Precio	NumFotos	NumBaños	Tamaño
28979149	9,1265	12500000,0000	11,0000	7,0000	2483,0000
31928760	37,9748	4900000,0000	5,0000	4,0000	7500,0000

COVratio: Valores menores que 0.93 o mayores que 1.07

	COVratio	Precio	NumFotos	NumBaños	Tamaño
28979149	0,0004	12500000,0000	11,0000	7,0000	2483,0000
29965257	1,2929	4100000,0000	1,0000	0,0000	4637,0000
33720107	1,1195	44000,0000	12,0000	12,0000	60,0000

Figura 110. Ejemplo de regresión 2. Observaciones influyentes. (Fuente: Elaboración propia).

6.6.5.5.4. Pestaña de Regresión 2: Observaciones influyentes 2

La pestaña que cierra el conjunto de opciones disponibles en la ventana *Regresión 2* es la denominada *Observaciones Influyentes 2*. Ésta permite al usuario la aplicación de los métodos propuestos en este trabajo para la identificación y la selección de observaciones influyentes, y al que se ha dedicado un capítulo previo.

Por este motivo, la primera selección que debe realizar el usuario es la del método a aplicar. El primero de los métodos es el de selección de observaciones influyentes mediante el método de tipificación. Recuérdese que el algoritmo construye, para cada observación de la muestra, un modelo de regresión con la submuestra resultante de eliminar dicha observación, y calcula su criterio de información. Para el vector de criterios de información de longitud n construido, se seleccionan los valores que son inferiores al valor medio de los criterios de información, menos k veces la desviación típica de éstos, donde k es una constante propuesta por el usuario, que toma el valor tres por defecto. El coste computacional de este método es elevado, por lo que el tiempo de cálculo se eleva considerablemente si el tamaño muestral es elevado.

Si computacionalmente es costoso el método anterior, el segundo método que puede ser seleccionado, lo es en mayor medida. Éste es el de selección de observaciones. En él, se construye el vector de criterios de información como en el método de detección de observaciones influyentes, se comprueba si hay algún valor que sea inferior al valor medio, menos k veces la desviación típica de éstos. Si no lo hay, el método finaliza. Si lo hay, se selecciona el menor de ellos, es decir, el que aumenta más significativamente la bondad de ajuste del modelo, y se elimina la observación correspondiente de la muestra. Este proceso se repite hasta que todos los valores de los criterios de información queden por debajo del límite definido.

El modelo final obtenido por el método de selección de observaciones se ajusta mejor, evidentemente, al conjunto de datos. Este método está justificado si el objetivo último del investigador es el de construir un modelo que ajuste bien los inmuebles con características más comunes en la muestra. No obstante, queda a criterio del investigador las observaciones que se dejarán fuera de la muestra y que pueden ser elegidas a través de la pestaña de *filtrado* en la ventana de *regresión*.

Como puede observarse en la Figura 111, el usuario debe seleccionar el criterio de información con el que desea aplicar el método. Como en una sección anterior del programa, se puede elegir entre el criterio de información de Akaike, el bayesiano o de Schwarz y otro criterio que utilice un factor de penalización, de las variables estimadas en el modelo, seleccionado por el usuario.

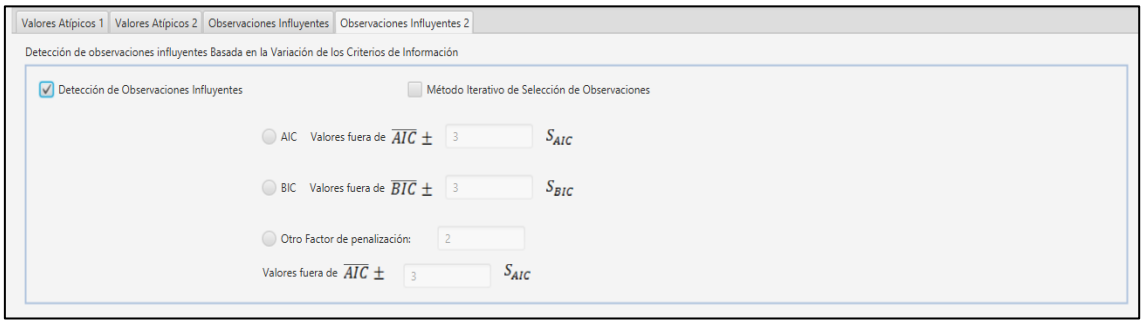


Figura 111. Regresión 2: Observaciones influyentes 2

El resultado del método de detección consta de una tabla en la que se muestran las observaciones consideradas influyentes, en la que se incluye el código del inmueble y los valores de todas las variables incluidas en el modelo, y un gráfico de barras en el que se representa, para cada observación, el valor del criterio de información del modelo asociado. Además, el título que precede la tabla indica el valor medio del vector del método y la frontera de influencia.

En esta tabla se incluyen también, si existen los valores que sean superiores a la media, k veces la desviación típica, que, aunque serían muy relevantes para la construcción del modelo, y, por tanto, tendrían una influencia opuesta en el modelo, pueden aportar información al usuario.

Aplicado este método al ejemplo comentado anteriormente, que está compuesto por un total de 480 locales a la venta de la ciudad de Sevilla, el resultado, que como puede observarse en la Figura 112, nos devuelve la existencia de dos inmuebles influyentes. El gráfico de barras, aunque difuso por el elevado número de observaciones de la muestra, muestra estos dos inmuebles, uno de ellos entre las observaciones cien y doscientos, y la segunda, poco antes de la observación trescientos.

El tiempo de cálculo fue muy reducido, inferior a los cuatro segundos.

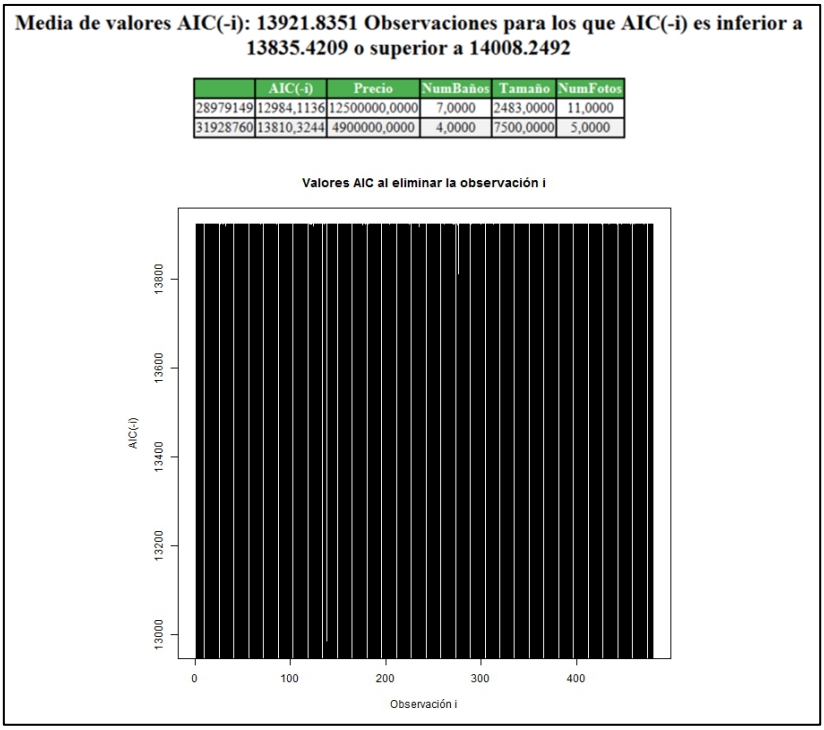
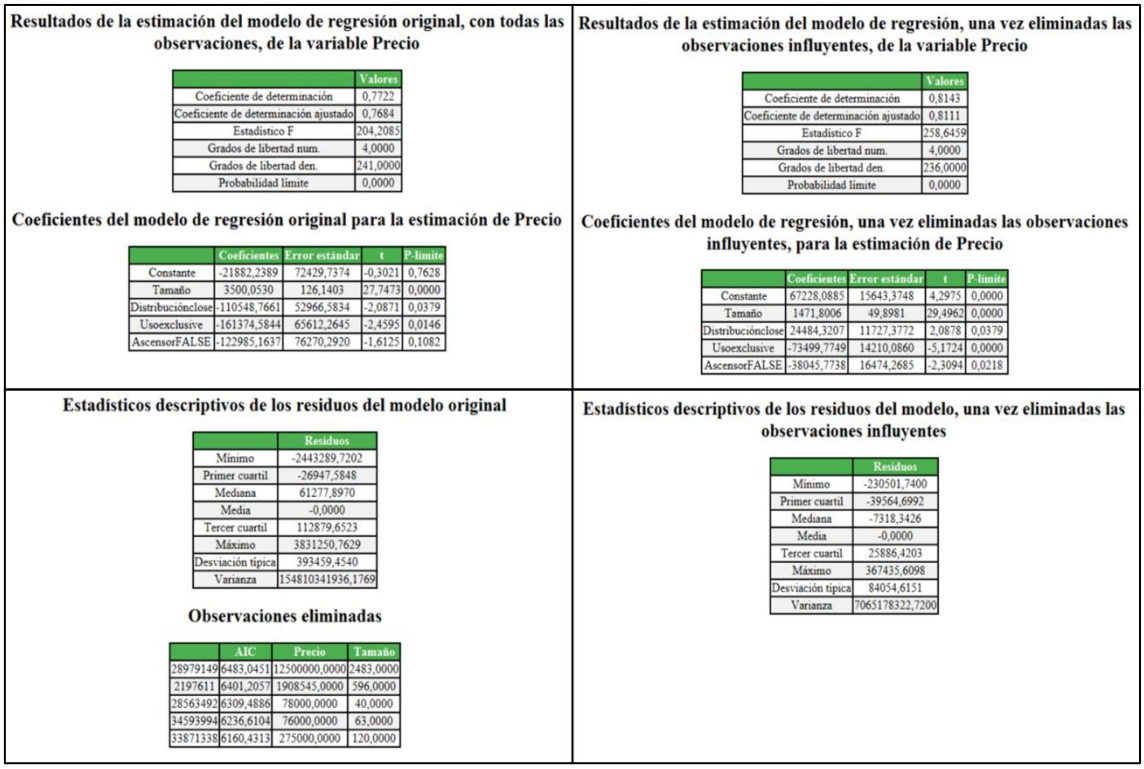


Figura 112. Ejemplo de regresión 2. Observaciones influyentes 2. Detección. (Fuente: Elaboración propia).

El resultado para el método de selección de observaciones se muestra en siete tablas. En las tres primeras se muestra el resultado del modelo estimado con todas las observaciones de la misma forma que se mostró en la ventana Regresión: en la primera tabla se incluyen los coeficientes de regresión y el contraste ANOVA de validación global, en la segunda la estimación de los coeficientes del modelo con sus respectivos contrastes de validación individual, y en la tercera un resumen descriptivo de los residuos del modelo. La cuarta tabla contiene la información de todas las observaciones eliminadas en el proceso, con los valores de AIC y los de sus variables en el modelo de regresión. Las tres últimas tablas contienen los resultados del modelo construido con las observaciones resultantes.

Se muestran, en la Figura 113, los resultados obtenidos para el método de selección, en un ejemplo que consta de oficinas a la venta en la ciudad de Sevilla, con un total de 246 inmuebles. Se ha optado por un modelo para la estimación del precio de la oficina a partir del tamaño, la distribución, el uso del edificio y la presencia de ascensor. Una vez aplicado el método, son eliminadas cinco oficinas. La duración del tiempo de ejecución fue ligeramente superior a los diez segundos.



Resultados de la estimación del modelo de regresión original, con todas las observaciones, de la variable Precio

	Valores
Coeficiente de determinación	0,7722
Coeficiente de determinación ajustado	0,7684
Estadístico F	204,2085
Grados de libertad num.	4,0000
Grados de libertad den.	241,0000
Probabilidad límite	0,0000

Coeficientes del modelo de regresión original para la estimación de Precio

	Coeficientes	Error estándar	t	P-límite
Constante	-21882,2389	72429,7374	-0,3021	0,7628
Tamaño	3500,0530	126,1403	27,7473	0,0000
Distribuciónclose	-110548,7661	52966,5834	-2,0871	0,0379
Usoexclusivse	-161374,5844	65612,2645	-2,4595	0,0146
AscensorFALSE	-122985,1637	76270,2920	-1,6125	0,1082

Resultados de la estimación del modelo de regresión, una vez eliminadas las observaciones influyentes, de la variable Precio

	Valores
Coeficiente de determinación	0,8143
Coeficiente de determinación ajustado	0,8111
Estadístico F	258,6459
Grados de libertad num.	4,0000
Grados de libertad den.	236,0000
Probabilidad límite	0,0000

Coeficientes del modelo de regresión, una vez eliminadas las observaciones influyentes, para la estimación de Precio

	Coeficientes	Error estándar	t	P-límite
Constante	67228,0885	15643,3748	4,2975	0,0000
Tamaño	1471,8006	49,8981	29,4962	0,0000
Distribuciónclose	24484,3207	11727,3772	2,0878	0,0379
Usoexclusivse	-73499,7749	14210,0860	-5,1724	0,0000
AscensorFALSE	-38045,7738	16474,2685	-2,3094	0,0218

Estadísticos descriptivos de los residuos del modelo original

	Residuos
Mínimo	-2443289,7202
Primer cuartil	-26947,5848
Mediana	61277,8970
Media	-0,0000
Tercer cuartil	112879,6523
Máximo	3831250,7629
Desviación típica	393459,4540
Varianza	1548103419 36,1769

Observaciones eliminadas

	AIC	Precio	Tamaño
28979149	6483,0451	12500000,0000	2483,0000
2197611	6401,2057	1908545,0000	596,0000
28563492	6309,4886	78000,0000	40,0000
34593994	6236,6104	76000,0000	63,0000
33871338	6160,4313	275000,0000	120,0000

Estadísticos descriptivos de los residuos del modelo, una vez eliminadas las observaciones influyentes

	Residuos
Mínimo	-230501,7400
Primer cuartil	-39564,6992
Mediana	-7318,3426
Media	-0,0000
Tercer cuartil	25886,4203
Máximo	367435,6098
Desviación típica	84054,6151
Varianza	7065178322,7200

Figura 113. Ejemplo de regresión 2. Observaciones influyentes 2. Selección. (Fuente: Elaboración propia).

6.6.5.6. La pestaña Redes Neuronales

Las pestañas anteriores utilizaban modelos de regresión para la estimación de una variable a partir de los valores de otras. En esta ocasión se ofrece la posibilidad de realizar la estimación a través de una red neuronal, el perceptrón multicapa, y compararla con la anterior.

Por tanto, está estructurada de la misma forma que las de regresión. Como puede observarse en la Figura 114, la parte superior de la ventana contiene las opciones necesarias para la selección de la variable dependiente que se desea estimar, las variables que se utilizarán para su estimación y la forma en la que éstas se relacionan. Una explicación detallada de su funcionamiento se ha realizado previamente en la pestaña de Regresión.

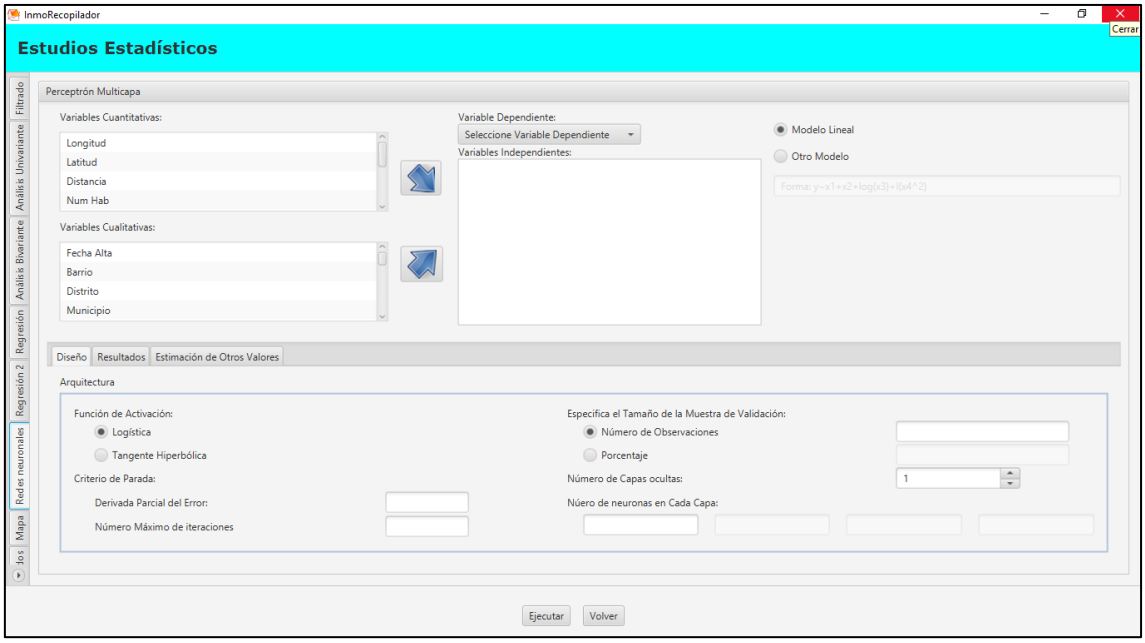


Figura 114. Pestaña de Redes neuronales

La introducción de variables cualitativas a través de variables *dummies* y la sintaxis necesaria para la construcción de un modelo distinto al estándar son iguales a los vistos anteriormente, y por tanto válidos para las pestañas *Regresión*, *Regresión 2* y *Redes Neuronales*.

La parte inferior de la ventana está compuesta a su vez de tres pestañas para la configuración de la red y la selección de resultados deseados. La dificultad en la modelización de la red y la aleatoriedad de las muestras utilizadas hacen que cada ejecución sea única, por lo que se hace necesario incluir la posibilidad de realizar estimaciones de forma simultánea a la construcción de la red.

No obstante, el usuario obtendrá entre otros resultados los valores de los pesos de la red resultante, por lo que podrá construir, a partir de la función de activación y con algo de paciencia la relación matemática hallada en la estimación.

Las tres pestañas disponibles, que detallaremos a continuación, son:

- Diseño.
- Resultado.
- Estimación de otros valores.

6.6.5.6.1. Pestaña de redes neuronales: Diseño

La pestaña Diseño, que puede verse también en la Figura 114, contiene todas las opciones necesarias para la construcción de la red de tipo perceptrón que se utilizará. Es necesario completar toda la información de la pestaña para la ejecución de la red neuronal.

En esta pestaña, el usuario deberá seleccionar la función de activación que se usará en la red, el tamaño de la muestra de validación, el criterio de parada del proceso de actualización de pesos, el número de capas ocultas de la red y el número de neuronas en cada capa.

La función de activación seleccionada por defecto es la función logística. Esta selección puede ser cambiada por la tangente hiperbólica. Al igual que se ha visto en pestañas anteriores, la selección de la muestra de validación se hace de forma aleatoria a partir del tamaño muestral seleccionado por el usuario aquí. También, en esta ocasión, se puede seleccionar el porcentaje de datos o el tamaño muestral que se reservará para medir la capacidad predictiva de la red.

El algoritmo de cálculo de los pesos de la red se detendrá siempre que se verifique uno de los dos criterios de parada que deben ser especificados por el usuario: La derivada parcial del error, que mide la variación de éste entre dos iteraciones consecutivas, y el número máximo de iteraciones que se realizarán.

El proceso terminará, por tanto, si la variación entre dos errores consecutivos es menor que la indicada por el usuario o si, por el contrario, aun no habiendo convergido la red, se llega al número máximo de iteraciones permitido. Esto último evita que el proceso continúe indefinidamente si el proceso llega a un mínimo local, y devolverá, como resultado, un mensaje indicando la divergencia de la red y recomendando la variación de los parámetros de la red.

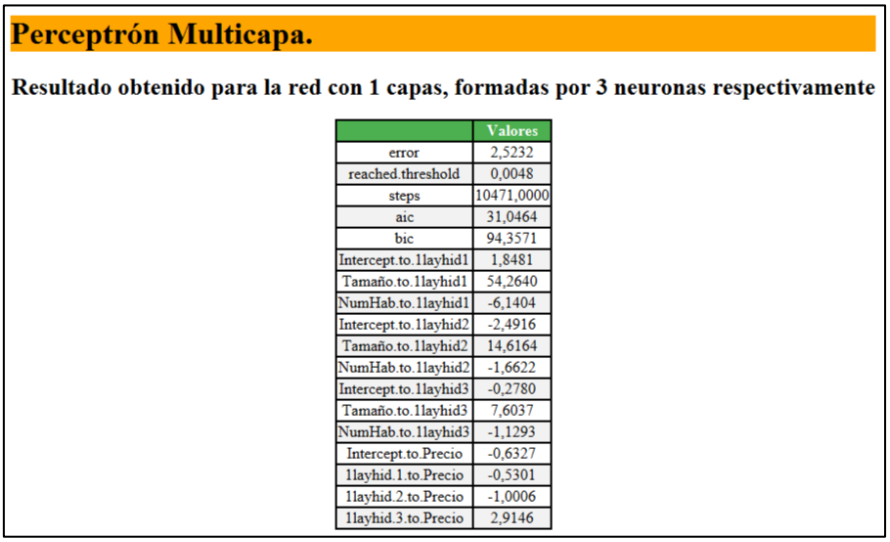
Las restantes opciones de la pestaña son las relativas a la arquitectura de la red. En concreto, estas opciones permiten el diseño de las capas ocultas de ésta, de manera que el usuario podrá seleccionar el número de capas ocultas que contendrá la red, entre una y cuatro, y el número de neuronas de cada una de las capas seleccionadas.

Es conveniente aclarar que, aunque se permite incluir en la red hasta cuatro capas ocultas, en condiciones normales, redes con una o dos capas ocultas son suficientes para obtener la mejor estimación posible.

El resultado de la red, una vez definidas todas las características de la red, sin necesidad de seleccionar ninguna otra opción, se expresa en una tabla en la que se indican:

- Error obtenido.
- Derivada parcial del error, que será inferior a la mínima exigida en la pestaña *Diseño*.
- Número de iteraciones realizadas.
- Valor del criterio de información de Akaike.
- Valor del criterio de información bayesiano o de Schwarz.
- Pesos de la red óptima.

El resultado para la estimación del precio, en una muestra de 1204 viviendas a la venta en la ciudad de Córdoba, a través de una red con función de activación logística, con una derivada parcial del error máxima de 0.005, un número máximo de iteraciones de 20000, una capa oculta, tres neuronas en dicha capa y una muestra de validación del 20 por ciento del total, se muestra en la Figura 115.



Perceptrón Multicapa.

Resultado obtenido para la red con 1 capas, formadas por 3 neuronas respectivamente

	Valores
error	2,5232
reached.threshold	0,0048
steps	10471,0000
aic	31,0464
bic	94,3571
Intercept.to.1layhid1	1,8481
Tamaño.to.1layhid1	54,2640
NumHab.to.1layhid1	-6,1404
Intercept.to.1layhid2	-2,4916
Tamaño.to.1layhid2	14,6164
NumHab.to.1layhid2	-1,6622
Intercept.to.1layhid3	-0,2780
Tamaño.to.1layhid3	7,6037
NumHab.to.1layhid3	-1,1293
Intercept.to.Precio	-0,6327
1layhid.1.to.Precio	-0,5301
1layhid.2.to.Precio	-1,0006
1layhid.3.to.Precio	2,9146

Figura 115. Ejemplo de redes neuronales. (Fuente: Elaboración propia).

6.6.5.6.2. Pestaña de redes neuronales: Resultados

Además de la tabla de resultados vista, están disponibles un conjunto de opciones que permiten ampliar la información ofrecida por la red. Estas opciones se configuran a través de la pestaña Resultados, que puede observarse en la Figura 116, y que son:

- El error cuadrático medio de la estimación de las observaciones de la muestra de validación. A esta opción se le puede añadir, además, el error cuadrático medio del modelo de regresión equivalente, de forma que se pueda comparar la bondad de los modelos construidos a través de los dos métodos. El resultado de ambos, se muestra en la misma tabla.
- Pesos iniciales de la red, que son asignados de forma aleatoria y de los que depende la convergencia de ésta. Los resultados se muestran en $k+1$ tablas, donde k es el número de capas ocultas de la red, de forma que en la tabla j se muestran los pesos iniciales de las neuronas de la red entre la capa $j-1$ y la capa j, donde denotamos la capa 0, como la capa que contiene las variables explicativas del modelo.
- Intervalos de confianza de los pesos de la red óptima, que sólo serán calculados si la red dispone, únicamente, de una capa oculta. Además de los extremos de los intervalos, se calcula el valor NIC o criterio de información de la red. El nivel de confianza de los intervalos será aportado también por el usuario, quién tendrá que indicarlo en valores sobre uno. El resultado de los intervalos se muestra con la misma estructura que la de los pesos iniciales, para cada extremo de los intervalos, es decir, cada tabla contendrá uno de los extremos de los intervalos de confianza de los pesos óptimos de la red entre la capa $j-1$ y j. En total, se mostrarán, por tanto, $2\cdot(k+1)$ tablas, y una última en la que se indica el valor del criterio de información de la red.
- El gráfico que representa la arquitectura de la red y que incluye los pesos obtenidos en la iteración óptima de la red.
- El gráfico de pesos generalizados respecto a cada una de las variables del modelo, que nos permite medir el grado de influencia que cada variable explicativa tiene sobre la variable dependiente. Este es un gráfico de dispersión en el que se representan en el eje de abscisas los valores normalizados de la variable explicativa y en el de ordenadas los valores de los pesos generalizados. Por tanto, se muestran tantos gráficos como variables explicativas contenga la red.

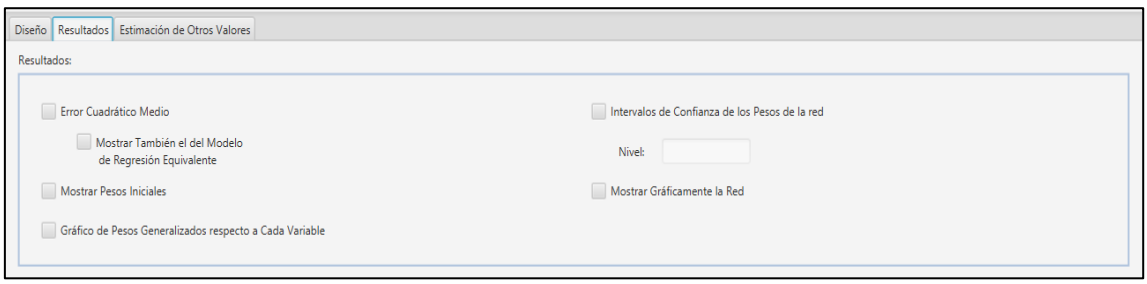


Figura 116. Redes neuronales: Resultados

Se han calculado los intervalos de confianza de los pesos de la red, obtenida para estimar el precio de venta, del ejemplo anterior, que constaba de 1204 viviendas de la ciudad de Córdoba, y en la que se utilizaron como variables explicativas el tamaño y el número de habitaciones de éstas. Los resultados se pueden observar en la Figura 117:

Error cuadrático medio para la estimación de Precio~ Tamaño+NumHab

	Valores
ECM para la red	6778794340,4132
ECM regresión	12566913849,9375

Pesos entre la capa 0 y 1

	N.1	N.2	N.3
AUX -> C1	0,3493	-0,5499	-0,9503
N.1-> C1	-0,6404	-0,2803	1,2924
N.2-> C1	0,0515	0,4959	-0,8251

Pesos entre la capa 1 y 2

	N.1
AUX -> C2	1,5973
N.1-> C2	-0,5960
N.2-> C2	-0,1404
N.3-> C2	-0,2924

Extremo inferior del intervalo de confianza de los pesos entre las neuronas de las capas 0 y 1

	N.1	N.2	N.3
AUX -> C1	1,0646	-0,2997	-2,8274
N.1-> C1	-5,4098	-53,9987	-5,2996
N.2-> C1	-1,0709	3,7359	-29,9295

Extremo inferior del intervalo de confianza de los pesos entre las neuronas de las capas 1 y 2

	N.1
AUX -> C2	0,7433
N.1-> C2	-1,6978
N.2-> C2	-0,4888
N.3-> C2	-1,2019

Extremo superior del intervalo de confianza de los pesos entre las neuronas de las capas 0 y 1

	N.1	N.2	N.3
AUX -> C1	3,3245	1,5969	0,2116
N.1-> C1	-0,9540	-22,4454	4,3614
N.2-> C1	1,1881	27,8248	-3,0777

Valor del criterio de información de la red

	Valor
NIC	2,6227

Extremo superior del intervalo de confianza de los pesos entre las neuronas de las capas 1 y 2

	N.1
AUX -> C2	1,8926
N.1-> C2	-0,3321
N.2-> C2	-0,2192
N.3-> C2	-0,2443

Figura 117. Ejemplo de redes neuronales. Intervalos y pesos. (Fuente: Elaboración propia).

También se ha representado gráficamente la arquitectura de la red y los pesos generalizados de ésta respecto a cada variable explicativa. La Figura 118, muestra el gráfico de la red y el de pesos generalizados para la variable Tamaño.

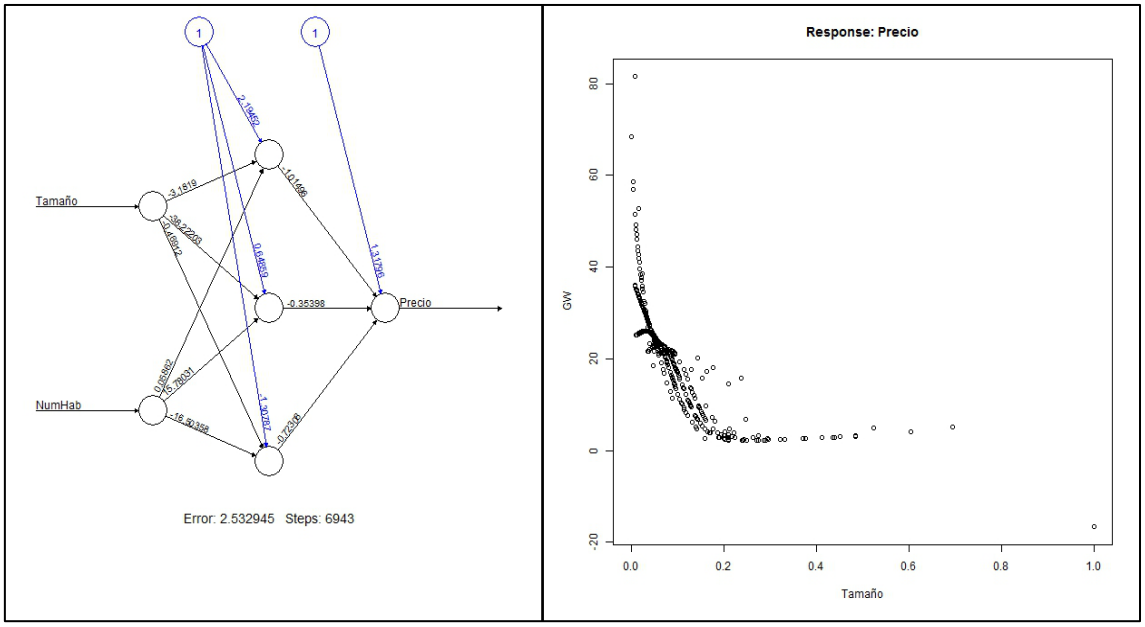


Figura 118. Ejemplo de redes neuronales. Gráficos. (Fuente: Elaboración propia).

6.6.5.6.3. Pestaña de redes neuronales: Estimación de otros valores

La tercera pestaña incluida en la ventana Redes Neuronales, *Estimación de Otros Valores*, nos permite incluir la estimación de valores de la variable dependiente en el resultado de la red. La aleatoriedad en la selección de los individuos que componen la muestra de validación, y que por tanto son eliminados de la muestra de entrenamiento, hacen prácticamente imposible la replicación exacta de los resultados de una red. Por ello es recomendable incluir, en las opciones de resultados, la estimación de los valores deseados en cada ejecución.

Como se muestra en la Figura 119, la introducción de los valores de las variables explicativas para los que se desea realizar la estimación, se realiza a través de una matriz que incluye en la columna de la izquierda, los nombres de éstas. En la parte superior derecha se puede observar un botón con el símbolo de suma. Cada vez que se hace clic sobre éste, se incluye una nueva columna a la tabla, es decir, una observación para la que se desea estimar la variable respuesta, cada una de las cuales toma por título *Observación*, precedida de un botón con un aspa. Al hacer clic sobre el aspa, se elimina la columna correspondiente.

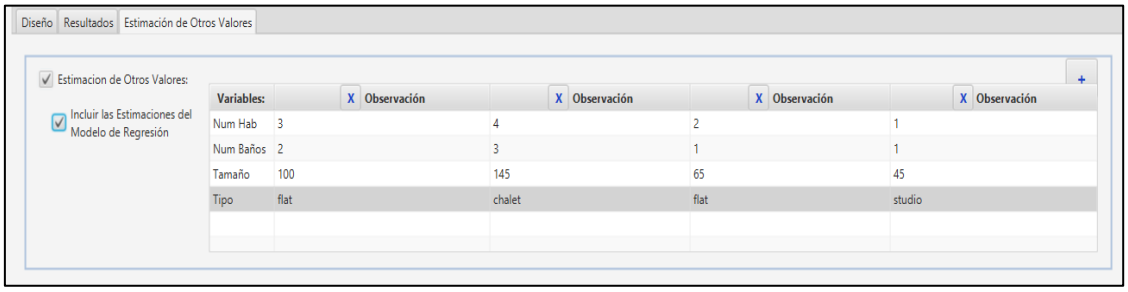


Figura 119. Redes neuronales: Estimación de otros valores

Para introducir valores para la predicción deberá pulsarse la tecla *Enter* tras escribir el valor deseado. Los valores de las variables cualitativas deberán escribirse de la misma forma en que están definidos, y deben ser valores incluidos en la muestra de estudio, por lo que no podrán incluirse modalidades que, aun siendo posibles en la definición de la variable cualitativa, no estén presentes en la muestra. Es decir, no es posible, por ejemplo, realizar la estimación del precio de una vivienda de tipo ático, si no existe ninguna vivienda de este tipo en la muestra de estudio.

También pueden incluirse las estimaciones dadas por el modelo de regresión equivalente para su comparación con las dadas por la red neuronal construida.

Los resultados de las predicciones solicitadas se muestran en una tabla que contiene, por filas, las predicciones dadas por la red, las obtenidas con el modelo de regresión si se han solicitado, así como los valores de las variables explicativas introducidas por el usuario para su identificación. Los valores de las variables cualitativas devueltos se corresponden con los de las variables *dummies* asociadas.

En la Figura 120 se muestra la tabla de resultados de las predicciones mostradas en la Figura 119 para la muestra de viviendas anteriormente comentada, en la que además se incluye el número de baños y el tipo de vivienda.

Predicciones de los valores dados

	Precio.estim.Red	Precio.est.Reg.	NumHab	NumBaños	Tamaño	Tipochalet	TipocountryHouse	Tipoduplex	Tipoflat	Tipopenthouse
1	172614,0242	187666,6727	3,0000	2,0000	100,0000	0,0000	0,0000	0,0000	1,0000	0,0000
2	276090,2507	285522,7239	4,0000	3,0000	145,0000	1,0000	0,0000	0,0000	0,0000	0,0000
3	94961,4665	114764,9411	2,0000	1,0000	65,0000	0,0000	0,0000	0,0000	1,0000	0,0000
4	39875,2505	89518,1750	1,0000	1,0000	45,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Figura 120. Ejemplo de redes neuronales. Predicciones. (Fuente: Elaboración propia).

6.6.5.7. La pestaña Mapa

Esta es la penúltima pestaña de la ventana de estudios estadísticos, y la última desde la que se puede solicitar algún estudio, ya que la última pestaña de esta ventana es en la que se muestran los resultados del estudio solicitado en el resto de pestañas.

La función principal de ésta, es la de ofrecer información geográfica de los inmuebles en estudio, y de la relación existente entre su localización y las características en estudio. Para ello se da como resultado una imagen satélite de la zona de estudio, en la que se superponen con puntos la localización de cada uno de los inmuebles. El color de los puntos vendrá dado por la característica en estudio y su valor para cada inmueble.

Si la variable es cualitativa, se definen tantos colores como modalidades tenga la variable y se colorean los puntos en función del color asignado a la modalidad. Esta

asignación será explicada mediante la leyenda correspondiente. Los colores utilizados en cada ejecución serán escogidos aleatoriamente entre la paleta disponible.

Si la variable es cuantitativa se presentará en la leyenda un degradado de colores desde el amarillo para los valores más pequeños, hasta el rojo para los valores más grandes. Las distintas tonalidades de naranja darán información sobre el menor o mayor valor de la variable.

Las opciones disponibles para la representación geográfica de los puntos se muestran en la ventana Mapa que puede observarse en la Figura 121.

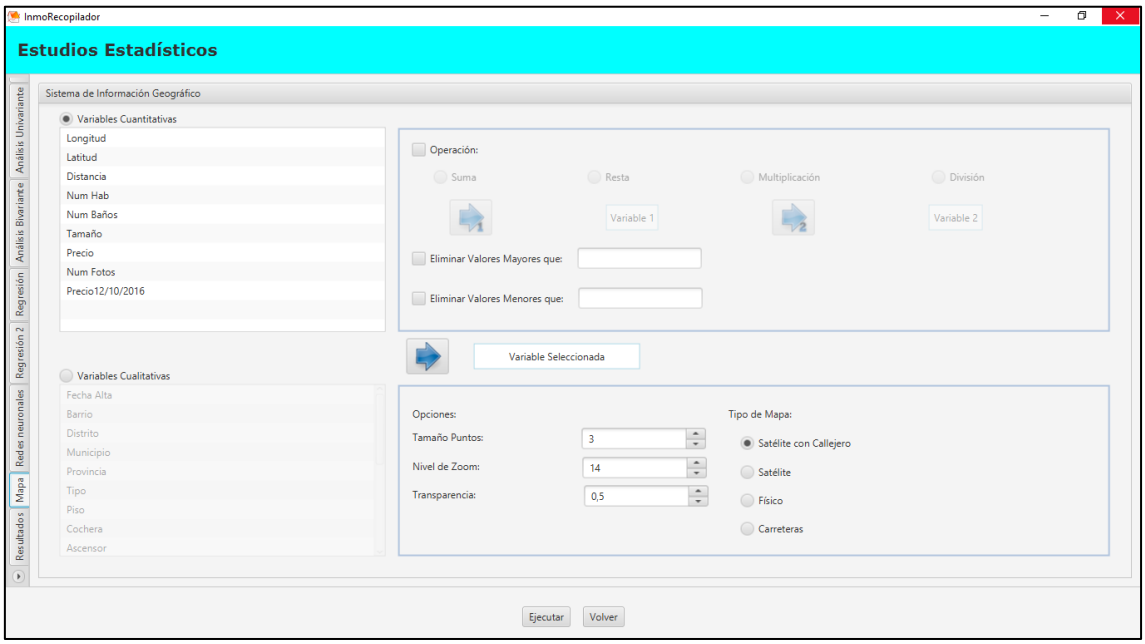


Figura 121. Pestaña Mapa

En primer lugar, como se ha comentado anteriormente, se seleccionará el tipo de variable que se desea representar a partir de las coordenadas geográficas de los inmuebles. En la parte superior izquierda tenemos el conjunto de variables cuantitativas disponibles y en debajo de éste, el listado de variables cualitativas. La selección de uno de los grupos excluye la posibilidad de utilizar una variable del otro grupo, por lo que es deshabilitado un listado al elegir el otro.

El lado derecho de la ventana está dividido en dos bloques. El bloque superior permite incluir, en la representación geográfica, una operación entre variables cuantitativas, y el acotamiento de los valores de los puntos que se van a representar. Este bloque sólo podrá

ser utilizado, por tanto, para la representación una característica cuantitativa de los inmuebles. El bloque inferior está constituido por el conjunto de opciones de la representación, y es común a todos.

La selección tanto de una variable cualitativa como de una única variable cuantitativa con la que no se desea operar, se hace haciendo clic sobre la variable elegida y pulsando el botón intermedio entre los dos bloques comentados en el párrafo anterior. No obstante, si se eligen dos variables cuantitativas y se desea representar una operación entre ellas, en primer lugar se deberá habilitar la opción Operación mediante el checkbox correspondiente, a continuación se seleccionará la operación entre las cuatro operaciones aritméticas permitidas y por último se elegirán las variables a operar, haciendo clic sobre la primera y pulsando el botón con la flecha numerado como uno y la segunda variable en la operación y pulsando en este caso el botón con la flecha numerado como dos.

Las últimas opciones de este bloque nos permiten, para variables cuantitativas, eliminar aquellos inmuebles con un valor mayor y/o menor que un valor determinado por el usuario. Para ello, basta con habilitar el checkbox correspondiente e introducir el límite deseado.

Las opciones disponibles para la representación gráfica son las siguientes:

- Tamaño de los puntos en el mapa. Se expresa a través de una escala en la que el tamaño de punto más pequeño disponible es uno, y el más grande diez con incrementos de 0,5. El valor por defecto es tres.
- Nivel de zoom: En esta opción se selecciona la escala del mapa que se mostrará. El usuario deberá ajustarla en función de la distancia máxima entre inmuebles, para no dejar ninguno fuera del ángulo de visión. Se pueden seleccionar valores naturales entre 10 y 18, de forma que el valor más pequeño se corresponde con una mayor escala. El valor por defecto es 14.
- Transparencia: Nivel de opacidad de los puntos en el mapa. Se seleccionará un valor entre cero, transparencia total de los puntos, y uno, máxima opacidad, con intervalos de cambio de 0.05. El valor por defecto es 0.5.
- Tipo de mapa. Los mapas mostrados son proporcionados por Google Inc., y se pueden seleccionar cuatro tipos de mapas:
    - Mapa satélite con callejero: Imagen satélite que además incluye el nombre de los barrios y de las calles más relevantes.
    - Mapa satélite: Imagen satelital que no incluye ninguna referencia de nombres.
    - Mapa físico: Imagen que muestra una vista de la planta de las calles, con los nombres más relevantes, e información física del terreno.

- Carreteras: En zonas urbanas este mapa es muy similar al anterior, sin información física. En zonas interurbanas, muestra además información del nombre de las principales vías de tráfico rodado.

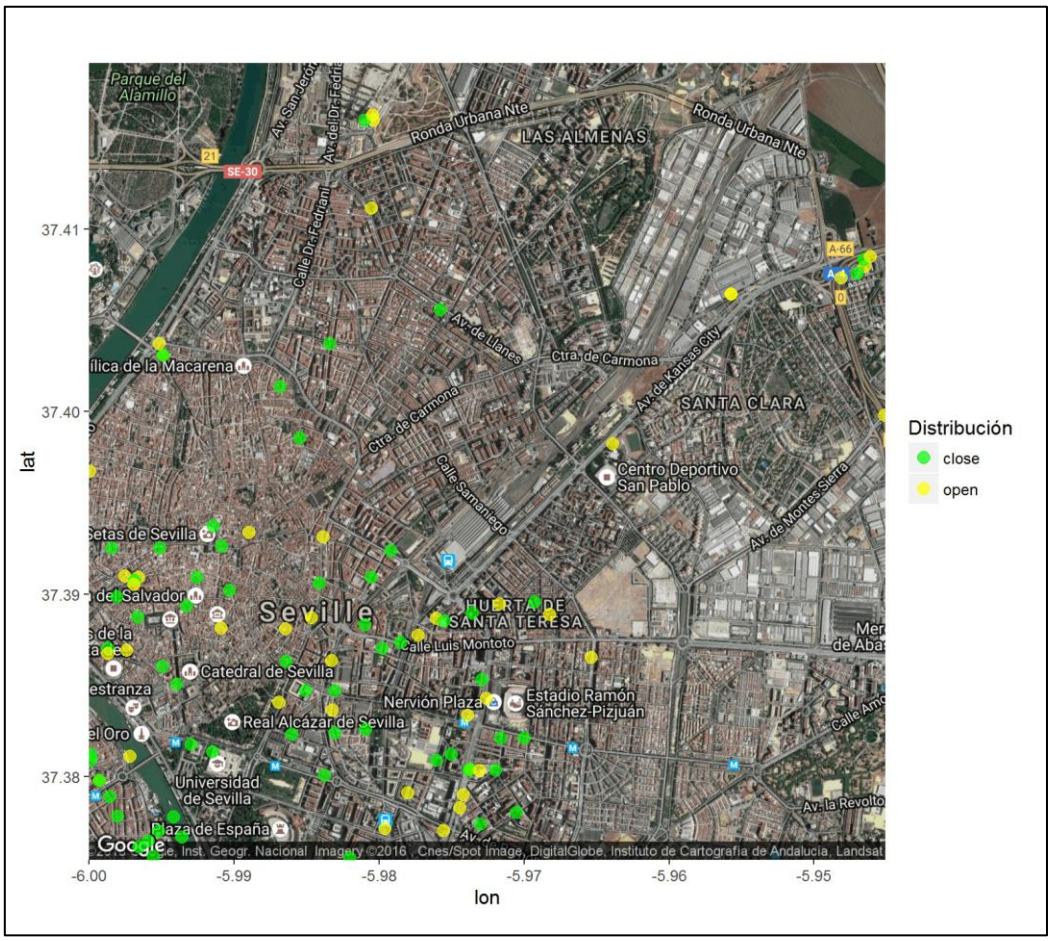


Figura 122. Ejemplo de mapa. (Fuente: Elaboración propia)

En la Figura 122, se muestra un mapa satélite con callejero de la ciudad de Sevilla en el que se sitúa una muestra de oficinas a la venta. El tamaño de los puntos es 3, el nivel de zoom es 14 y la transparencia 0.7. Con los diferentes colores se indica la distribución del recinto en el que se sitúan las oficinas, de forma que con color verde se indica que la oficina tiene una distribución tabicada en varias estancias, y con color amarillos las oficinas diáfanas.

6.6.5.8. La pestaña Resultados

Todos los resultados obtenidos a partir de los estudios estadísticos son impresos en la ventana de Resultados. Cada estudio es precedido de un título sobre fondo naranja con el

nombre del estudio. Cada vez que un nuevo estudio es solicitado, éste se inicia con un nuevo título, seguido del conjunto de tablas, y finaliza con los gráficos solicitados.

Como puede observarse en la Figura 123, la ventana contiene, además del área de resultados, dos botones en la parte superior derecha. El primero de ellos permite guardar en disco un archivo HTML con todos los resultados. El segundo, limpia la pantalla eliminando todas las tablas y gráficos.

Es importante destacar que si se cierra la ventana de estudios estadísticos también se limpiará la ventana de resultados, por lo que si se desea conservar las tablas y gráficos obtenidos deberá guardarse el archivo antes del cierre.

Figura 123. Pestaña Resultados

Al hacer clic sobre el botón de guardado del archivo, se abre un cuadro de diálogo como el que se muestra en la Figura 124, en el que el usuario debe indicar el nombre de archivo deseado y la carpeta de destino del mismo.

Al guardar el archivo se generará también una carpeta con el mismo nombre que el archivo, seguido de un guion bajo y de la palabra *files*.

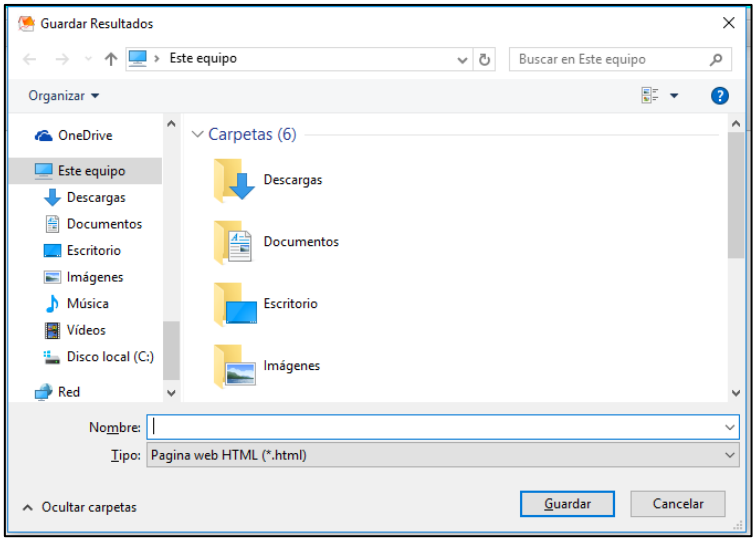


Figura 124. Pestaña Resultados. Guardar

Los resultados obtenidos pueden ser copiados y pegados en un procesador de textos. Basta con seleccionar la tabla y pulsar *Ctrl + C*. Los gráficos pueden guardarse en el portapapeles de Windows haciendo clic con el botón derecho del ratón y seleccionando la opción *Copiar Imagen en Portapapeles*.

7.Especificación de las funciones implementadas en lenguaje R

Para el funcionamiento del programa, se han programado 8 funciones en R con aproximadamente 3500 líneas de código, y en las que se hace uso de algunas funciones ya implementadas en paquetes por otros equipos, que son de libre uso y que serán indicadas a continuación. Las funciones implementadas son:

- Resumen.
- Explora.
- Contingencia.
- Correlacion.
- Regresion.
- Regresion2.
- Perceptron.
- Mapa.

A continuación, se describen cada una de estas funciones, explicando su funcionamiento y los paquetes R utilizados.

7.1. Función Resumen

Esta función se encuentra en el archivo *resumen.R*. Se ejecuta desde el programa siempre que el usuario desee realizar un estudio descriptivo de una o más variables cuantitativas como son: *Longitud*, *Latitud*, *Distancia*, *NumHab*, *NumBaños*, *Tamaño*, *Precio*, *NumFotos* y las variables correspondientes a las distintas actualizaciones de precios. Sus argumentos son los siguientes:

- Un vector lógico que denominamos *opcionesL*, y que incluye la información sobre la selección realizada por el por el usuario en la pestaña Análisis Univariante de la ventana Estudios estadísticos. Se compone de un total de 19 elementos.
- Un vector de elementos numéricos, denominado *percentiles*, que contiene los percentiles seleccionados por el usuario para ser calculados.
- Una tabla de datos, que recibe R en formato de lista, que contiene los datos correspondientes a las variables cuantitativas que se desean analizar.
- La variable numérica *ncon* que contiene un valor numérico entre 0 y 1 con el nivel de confianza con el que el usuario desea construir el intervalo de confianza para el valor medio de las variables seleccionadas.
- Un vector lógico, de longitud 5, denominado *grafL*, que transmite la información acerca de los gráficos seleccionados por el usuario para ser representados.
- Por último, la variable de tipo cadena que incluye la información del directorio en el que se guardan las librerías de R necesarias para la correcta ejecución del programa.

La función resumen devuelve una lista formada por dos elementos, el primero de tipo matriz, en la que incluyen en forma tabular, los resultados de los estudios solicitados por el usuario, y el segundo, un vector con elementos de tipo cadena con los nombres de los archivos gráficos generados en una carpeta del disco duro, y desde donde serán leídos para ser mostrados en la ventana de resultados.

En primer lugar, se extraen de la lista los valores de las variables de estudio para poder ser utilizados. A continuación, se crea el vector de nombres de la matriz de resultados, en el que se indicará el nombre de cada estadístico calculado, en función de la selección realizada por el usuario.

A partir de este momento, se calculan los valores de las medidas solicitadas de todas las variables, y los resultados se incluyen como fila en la matriz de resultados.

Las medidas calculadas a través de la función resumen son:

- Medidas de posición: Media, mediana y percentiles.
- Medidas de dispersión: Desviación típica muestral, varianza muestral, mínimo, máximo, rango o recorrido y coeficiente de variación.
- Medidas de forma: Coeficiente de asimetría y coeficiente de apuntamiento o *curtosis*. Para éstas medidas se ha utilizado la librería *e1071*, incluida en el repositorio CRAN, que no es cargada en la instalación típica del programa, y que es obra de Meyer, Dimitriadou, Hornik, Weingessel, y Leisch, (2015)
- Extremos del intervalo de confianza para la media, para un nivel de confianza dado por *ncon*. Estos sólo serán calculados si la varianza del número de datos es distinta de 0.
- Contrastes de normalidad de las variables seleccionadas. La función devuelve en los tres casos el valor del estadístico de contraste, y el de la probabilidad límite de éste. Se pueden realizar tres contrastes de hipótesis:
    - El test de *Shapiro – Wilks* que sólo admite tamaños de muestra de hasta 5000 datos, por lo que en caso de que este límite sea superado, el programa extrae una muestra aleatoria de tamaño 5000 y le aplica el test solicitado.
    - El test de *Lilliefors,* – conocido como de *Kolmogorov – Smirnov* - con la función *lillie.test* del paquete *nortest*, creado por Gross y Ligges, (2012), con licencia GPL igual o mayor de 2, y también incluido en el repositorio CRAN de la comunidad R.
    - El test de *Jarque – Bera*, que se ha construido apoyado en el uso de los valores dados por el paquete *e1071* para los coeficientes de asimetría y apuntamiento, antes mencionado. En caso de valores faltantes, estos se han eliminado.

Para representar gráficamente los datos, el usuario podrá elegir uno o más, de los siguientes: gráfico de barras, gráfico de sectores, histograma y diagrama de caja. Esta información es trasladada a la función a través del vector lógico *grafL*, con los valores de los primeros cuatro elementos indicando si cada uno de estos cuatro gráficos es seleccionado. El usuario puede también elegir si los gráficos resultantes deben realizarse a partir de las frecuencias o por el contrario deben ser representados a partir de los porcentajes de los datos dados. Esta información es trasladada a través del quinto elemento del vector *grafL*.

Cabe realizar tres consideraciones acerca de los gráficos representados:

- El histograma de frecuencias se representa incluyendo la campana de Gauss correspondiente a la representación de la función de densidad en el supuesto de normalidad de la variable representada.
- El número de intervalos en el que se divide el rango viene dado por la fórmula dada por Sturges, (1926), en la que determina que el número de intervalos óptimo para un conjunto de n datos es:
- $N^{\underline{o}}\ intervalos = 1 + 3.322 \cdot \log n$

- Se representa un gráfico para cada tipo y variable seleccionados. Estos gráficos son guardados en una carpeta personal con nombres que indican el tipo y variable implicada, seguido de un valor aleatorio, que impide que en sucesivas ejecuciones del programa sean sobrescritos. Los nombres de los gráficos construidos, en formato *jpeg*, se guardan en un vector denominado *graficos*, que se incluye como segundo elemento de la lista, para que el programa identifique los gráficos que debe representar en cada ejecución. Una vez que el usuario cierra el programa, estos archivos son eliminados.

7.2. Función Explora

Esta función se encuentra en el archivo *exploratorio.R*. Se ejecuta a través, también, de la pestaña Análisis Univariante de la ventana Estudios estadísticos, pero en este caso, cuando el usuario desee realizar un estudio descriptivo de una o más variables cuantitativas en función de alguno de las variables cualitativas de que se disponen. Las variables cuantitativas disponibles son: *Longitud*, *Latitud*, *Distancia*, *NumHab*, *NumBaños*, *Tamaño*, *Precio*, *NumFotos* y las variables correspondientes a las distintas actualizaciones de precios. Los factores disponibles son: *Dirección*, *Barrio*, *Distrito*, *Municipio*, *Provincia*, *Tipo*, *Piso*, *Cochera*, *Ascensor*, *Piscina*, *AireAcon*, *Terraza*, *Trastero*, *Tendederos*, *Empotrados*, *Estado*, *Localización*, *Esquina*, *Salida de Humos*, *Seguridad*, *Distribución*, *Uso del Edificio*, *Agua Caliente Independiente*, *Calefacción independiente*, *Puerta Automática y Parking de Motocicletas*, en función del tipo de inmueble estudiado. Sus argumentos son los siguientes:

- Un vector lógico que denominamos *opcionesL*, y que incluye la información sobre la selección realizada por el por el usuario en la pestaña Análisis Univariante de la ventana Estudios estadísticos. Se compone de un total de 19 elementos.
- Un vector de elementos numéricos, denominados percentiles, que contiene los percentiles seleccionados por el usuario en caso de querer calcularlos.
- Una tabla de datos denominada casas, que recibe R en formato de lista, y que contiene los datos correspondientes a las variables cuantitativas que se desean analizar.
- Una tabla de datos denominada factores, que al igual que casas, recibe R en formato de lista. Contiene los datos correspondientes a las variables cualitativas o factores que se desean utilizar para clasificar las distintas variables cuantitativas.
- La variable numérica *ncon* que contiene un valor numérico entre 0 y 1 con el nivel de confianza con el que el usuario desea construir el intervalo de confianza para el valor medio de las variables seleccionadas.
- Un vector lógico, de longitud 5, denominado *grafL*, que transmite la información acerca de los gráficos seleccionados por el usuario para ser representados.

- Por último, la variable de tipo cadena que incluye la información del directorio en el que se guardan las librerías de R necesarias para la correcta ejecución del programa.

La función Explora devuelve una lista formada por $k \cdot f + 1$ elementos, donde k es el número de variables cuantitativas seleccionadas, y f el número de factores de clasificación. Es decir, para cada combinación de parejas variable – factor, se genera un elemento de la lista que será de tipo matriz. El número de filas de la matriz dependerá con las opciones de estudio seleccionadas y el nombre de cada fila indicará el estadístico calculado en cada caso. El número de columnas será igual al número de categorías del factor considerado en cada caso.

El último elemento, al igual que en el resto de funciones, está formado por el vector de caracteres que incluye el nombre de los gráficos generados en la petición del usuario.

En primer lugar, se extraen de las listas los dos conjuntos de datos recibidos, casas y factores, y se definen como variables cuantitativas y cualitativas respectivamente. A continuación, se construye, a partir del vector *opcionesL*, el vector de caracteres que formará la lista de nombres de todas las matrices generadas, y que expresan el estadístico calculado en cada caso.

Una vez realizado esto, se calculan las medidas solicitadas por el usuario, según los valores del vector lógico *opcionesL*, tal y como se ha descrito para la función resumen. Esto se hace para cada una de las variables cuantitativas seleccionadas en función de las distintas categorías de cada uno de los factores seleccionados.

Al dividir la muestra en grupos dados por las distintas categorías de los factores, es necesario realizar controles sobre el tamaño muestral resultante en cada grupo y sobre la no nulidad de la varianza de éste lo que devolvería error de ejecución al calcular el intervalo de confianza para la media o los contrastes de normalidad. En los casos en los que no es posible su cálculo, el sistema devuelve un valor NA en el elemento de la matriz - resultado correspondiente.

Al igual que en la función resumen, la limitación impuesta por la función *shapiro.test* para muestras de hasta tamaño 5000 se controla y se resuelve, en caso de incumplirse, seleccionando una muestra aleatoria de tamaño 5000.

Los paquetes no incluidos en R que se han utilizado en esta función son e1071 implementado por Meyer et al. (2015) para el cálculo de los coeficientes de forma; *nortest* cuyos autores son Gross y Ligges (2012); y en este caso, se ha utilizado el paquete *tseries*, desarrollado por Trapletti y Hornik (2017), que puede descargarse desde el repositorio *CRAN* de R, y cuya licencia de uso es *GPL – 2*, para el cálculo del estadístico y la probabilidad límite dada por el test de *Jarque – Bera*. La función utilizada es *jarque.bera.test*.

Al igual que en la función resumen, el usuario puede representar gráficamente los datos mediante gráficos de barras, gráficos de sectores, histogramas y diagramas de caja. Además, puede escoger entre representar los valores de frecuencias o los porcentajes correspondientes. En este caso, para cada combinación de variable cuantitativa con factor, se realizan tantas representaciones gráficas como categorías tenga el factor. Se representan todos los gráficos de una misma combinación en el mismo archivo, dividiendo el área de representación en tantas partes como es necesario. No obstante, por las limitaciones impuestas por el software, no es posible realizar la representación conjunta cuando el factor está compuesto por más de nueve categorías. Cuando esto ocurre, se representan los gráficos de manera individual.

El histograma se representa junto a la campana de Gauss que muestra la curva dada por su función de densidad bajo el supuesto de normalidad. Además, el número de rectángulos definidos en este diagrama viene dado por la fórmula de Sturges (1926).

7.3. Función Contingencia

Esta función se encuentra en el archivo *contingencia.R*. Su ejecución tiene lugar a través del programa desde la pestaña Análisis Bivariante de la ventana Estudios estadísticos, cuando el usuario desea analizar el grado de asociación entre variables cualitativas o factores, a través del estudio de sus tablas de contingencia. Las variables disponibles son: *Dirección, Barrio, Distrito, Municipio, Provincia, Tipo, Piso, Cochera, Ascensor, Piscina, AireAcon, Terraza, Trastero, Tendederos, Empotrados, Estado, Localización, Esquina, Salida de Humos, Seguridad, Distribución, Uso del Edificio, Agua Caliente Independiente, Calefacción independiente, Puerta Automática y Parking de Motocicletas,* en función del tipo de inmueble estudiado. El estudio se realiza para cada

combinación de dos de las variables elegidas por el usuario. Sus argumentos son los siguientes:

- Un vector lógico que denominamos *opcionesL*, y que incluye la información sobre los estudios activados por el usuario en la pestaña Análisis Bivariante de la ventana Estudios estadísticos. Se compone de un total de 9 elementos.
- La tabla de datos denominada factores, y recibida en formato lista desde el programa matriz, que contiene las variables, en columna, para las que se quiere realizar el estudio solicitado.
- La variable ruta, de tipo cadena, que incluye la ruta donde R debe buscar los paquetes necesarios para la correcta ejecución del programa. No obstante, en esta función no se han utilizado paquetes adicionales a los instalados por R en la instalación típica, por lo que no es necesaria la existencia de este argumento. Se ha incluido por homogeneización con el resto de funciones.

Si el número de variables cualitativas seleccionadas por el usuario es k, se analizan un total de $\binom{k}{2}$ parejas de variables, para determinar la existencia de asociación entre ellas. Al igual que en el resto de funciones, la función Contingencia devuelve una lista formada por $\binom{k}{2} \cdot m + 1$ elementos, donde m es el número de elementos seleccionados para el estudio, excluida la opción gráfica. El último elemento de la lista es el vector que contiene los nombres de los archivos gráficos generados en la ejecución.

Una vez extraído el conjunto de datos de la lista recibida y definidos como factores, el estudio realizado por la función Contingencia, según los valores lógicos del vector *opcionesL* es el siguiente:

- Tabla de contingencia de frecuencias observadas entre ambos factores, que incluye las frecuencias marginales de ambas variables.
- Tabla de contingencia de frecuencias esperadas, incluyendo también las frecuencias marginales.
- Tabla de contingencia de porcentajes sobre el total de la muestra con distribución marginal de porcentajes.
- Tabla de contingencia de porcentajes sobre el total de cada fila de la tabla, es decir, tomando como distribución total la de cada categoría de la variable expresada por filas en la tabla.
- Tabla de contingencia de porcentajes sobre el total de cada columna de la tabla.
- Tabla de contingencia de residuos estandarizados corregidos. Estos valores permiten medir el grado de asociación de las variables. Su fórmula es:

$$e_{ij}^{*} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij} \cdot \left(1 - {}^{n_{i\cdot}}/_{n}\right) \cdot \left(1 - {}^{n_{\cdot j}}/_{n}\right)}}$$

Donde n_{ij} es la frecuencia observada, e_{ij} la frecuencia esperada, $n_{i\cdot}$ es la frecuencia marginal de fila i, $n_{\cdot j}$ es la frecuencia marginal de la columna j, y n es el tamaño muestral.

- Contraste *Chi – cuadrado* de independencia para tablas de contingencia. El estadístico de contraste es:

$$\chi^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

  En tablas con dos filas y dos columnas se utiliza la corrección por continuidad. En caso contrario, se aplica el test Monte Carlo con 2000 replicaciones, dado por Hope (1968).

- Test exacto de *Fisher* para independencia.

  En tablas 2 x 2 se utiliza el contraste de Fisher (1935), que se basa en que en el caso de independencia, para unas frecuencias marginales dadas, el primer elemento de la tabla seguirá una distribución *hipergeométrica* no centrada, con parámetros dados por los *odds ratios*.

  En tablas de orden superior, el cálculo de la probabilidad límite se calcula a través de la red desarrollada por Mehta y Patel (1986) para *FORTRAN*, y posteriormente mejorado por Clarkson, Fan, y Joe (1993).

- Por último, se representa un gráfico de barras agrupadas, en el que se representan las frecuencias de las categorías de una de las variables, dividiendo en distintas barras en función de las categorías de la otra variable.

7.4. Función Correlacion

Esta función se encuentra en el archivo *correlacion.R*. Se invoca desde el programa principal a través de la misma ventana que la función anterior, la pestaña Análisis Bivariante de la ventana Estudios estadísticos. No obstante, ésta se activa cuando el usuario selecciona la opción "*variables cuantitativas*", es decir, cuando se desea analizar la existencia de relación lineal entre variables cuantitativas y su grado. Las variables disponibles son: *Longitud, Latitud, Distancia, NumHab, NumBaños, Tamaño, Precio, NumFotos* y las correspondientes a las distintas actualizaciones de los precios de un inmueble. Los argumentos de la función son:

- El vector *opcionesL*, de tipo lógico, compuesto por siete elementos con los estudios seleccionadas por el usuario para incluir en el resultado de la función.

- Una tabla de datos en formato lista, denominada *casas*, con los datos de las variables para las que se desea realizar el estudio de correlación.
- Una variable de tipo cadena con la ruta en la que se encuentran los paquetes adicionales necesarios para la ejecución de la función.

Para el estudio de la relación entre las variables seleccionadas se calcula tanto el correspondiente coeficiente para la información muestral dada, como el contraste de hipótesis asociado para determinar la existencia o no de relación lineal a nivel poblacional. Los resultados que pueden obtenerse son:

- Matriz de coeficientes de correlación lineal de *Pearson* para cada par de variables seleccionadas. Se calcula:

$$r_{jk} = \frac{n \cdot \sum_{i=1}^{n} x_{j_i} \cdot x_{k_i} - \sum_{i=1}^{n} x_{j_i} \cdot \sum_{i=1}^{n} x_{k_i}}{\sqrt{n \cdot \sum_{i=1}^{n} {x_{j_i}}^2 - \left(\sum_{i=1}^{n} x_{j_i}\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} {x_{k_i}}^2 - \left(\sum_{i=1}^{n} x_{k_i}\right)^2}}$$

- Así como su estadístico de contraste, que verifica, bajo hipótesis de normalidad bivariante:

$$t = \frac{r_{jk}}{\sqrt{\frac{1 - r_{jk}^2}{n-2}}} \sim t(n-2)$$

- Matriz de coeficientes de correlación de *Spearman*, que se calculan de la siguiente forma:

$$r_s = 1 - \frac{6}{n^3 - n} \cdot \sum_{i=1}^{n} d_i^2$$

- Donde d_i son las diferencias de los rangos de cada par de variables X_j y X_k para $i = 1, \dots, n$. Por tanto, el coeficiente de correlación de *Spearman* no es más que el coeficiente de correlación de Pearson de los rangos de las variables. También se calcula la matriz de estadísticos y las de las probabilidades límite de los contrastes de significatividad de los coeficientes.
- Para muestras suficientemente grandes, se verifica que:

$$t = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1 - r_s^2}} \approx t(n-2)$$

- Matriz de coeficientes de correlación por rangos de *Kendall*. Para el cálculo se utiliza el paquete *Kendall*, creado por McLeod, (2005) con licencia GPL que puede descargarse del repositorio *CRAN*. El coeficiente de correlación de *Kendall* suele usarse con frecuencia para el caso de que se den una gran cantidad de empates. Su fórmula es:

$$\tau = \frac{2 \cdot (N_C - N_D)}{\sqrt{n \cdot (n-1) - T_i} \cdot \sqrt{n \cdot (n-1) - T_j}}$$

- Donde N_C y N_D son los números de pares concordantes y discordantes respectivamente, T_i y T_j el número de empates en las variables X_i y X_j respectivamente, y n el tamaño muestral. El estadístico τ converge rápidamente a una distribución Normal con media 0 y desviación típica:

$$\sigma_\tau = \frac{1}{3} \cdot \sqrt{\frac{4n + 10}{n \cdot (n-1)}}$$

- A continuación, se calculan los coeficientes de correlación parcial de Pearson, *Spearman* y *Kendall*, es decir, se analiza la existencia de relación entre cada dos variables, eliminando la posible influencia del resto de variables consideradas. Para el cálculo de estas medidas se ha utilizado la librería *ppcor*, implementada por Kim, (2015). Estos se calculan a partir de los valores de correlación simple, y su formulación depende del número de variables implicadas.

7.5. Función Regresion

Esta función se encuentra en el archivo *regresion.R*. Una vez seleccionadas por el usuario la variable dependiente del modelo, las variables independientes, la formulación del modelo, y alguna de las opciones disponibles se ejecuta la función *Regresion* través de la pestaña Regresión de la ventana Estudios estadísticos.

Como variable dependiente, el usuario puede elegir entre alguna de las variables cuantitativas que son: *Longitud, Latitud, Distancia, NumHab, NumBaños, Tamaño, Precio, NumFotos* y las correspondientes a las distintas actualizaciones de los precios de un inmueble.

Las variables que pueden ser seleccionadas como independientes en el modelo de regresión son cualquiera de las variables cuantitativas indicadas anteriormente, salvo la seleccionada como independiente, y cualquier variable cualitativa disponible según el tipo de inmueble que se analice. Estas son: *Dirección, Barrio, Distrito, Municipio, Provincia, Tipo, Piso, Cochera, Ascensor, Piscina, AireAcon, Terraza, Trastero,*

Tendederos, Empotrados, Estado, Localización, Esquina, Salida de Humos, Seguridad, Distribución, Uso del Edificio, Agua Caliente Independiente, Calefacción independiente, Puerta Automática y Parking de Motocicletas, en función del tipo de inmueble estudiado.

Cuando se incluyen en el modelo de regresión una o más variables cualitativas, la función las incluye en éste, generando variables auxiliares o *Dummies*. Para cada variable cualitativa, se generan tantas variables *dummies* como número de categorías tenga ésta menos una, para evitar así problemas de multicolinealidad exacta. Cada variable auxiliar generada tomará el nombre de la variable cualitativa de la que procede seguido del nombre de la categoría a la que está asociada, tomando el valor 1 para ésta, y 0 para el resto.

Así, por ejemplo, la variable Piscina puede tomar dos valores, *FALSE* y *TRUE*, por lo que la función generará una variable *dummy* denominada *PiscinaFALSE*, que tomará el valor 1 para las viviendas que carezcan de piscina y 0 para las que sí tengan.

Los argumentos requeridos por la función son:

- Un vector lógico de longitud 33, denominado *opcionesL*, que contiene las especificaciones seleccionadas por el usuario, y que indicarán el conjunto de resultados que ofrecer y el método para el cálculo de éstos.
- Un vector numérico que está compuesto por los valores de la variable dependiente. Este vector se recibe en el entorno de R como una lista, por lo que, en primer lugar, se extrae.
- Una tabla de datos denominada *independV*, que contiene, también en formato lista, los valores de las variables cuantitativas seleccionadas como variables independientes del modelo.
- Una tabla de datos de tipo cadena, con formato lista, denominada *independF*, que contiene los valores originales de las variables cualitativas seleccionadas para formar parte del modelo de regresión. A partir de éstas se construye el conjunto de variables *dummies*, como se ha explicado anteriormente.
- Un elemento de tipo cadena denominado *form*, que contiene la formulación del modelo de regresión seleccionada por el usuario en caso de no seleccionar el modelo lineal básico que se desprende de las variables seleccionadas. Este elemento deberá estar correctamente redactado según se especifica en el manual de uso del programa.
- Un valor real positivo, denominado pen, que incluye el factor de penalización seleccionado por el usuario para el cálculo del criterio de información (dos para el criterio de información de *Akaike* y logaritmo del número de observaciones para el de *Schwarz*).

- La variable muestra, que contiene un número real en el que el usuario indica el número de observaciones que desea omitir en la estimación del modelo para poder usarlas posteriormente en el cálculo de las medidas de bondad de predicción. Es decir, se divide la muestra en dos: la muestra de estudio o estimación y la muestra de validación. Además, esta variable puede indicar el valor absoluto del tamaño muestral para la validación, o, por el contrario, la proporción de la muestra usada.
- *Nivel*: Es un valor numérico entre cero y uno, que contiene el nivel de confianza con el que se construirán los intervalos de confianza para la estimación de los coeficientes del modelo de regresión.
- Al igual que *form*, la variable *form2* es la expresión de un modelo de regresión, construido a partir de las variables seleccionadas. Ésta será recibida siempre que el usuario desee aplicar el contraste de *Davidson* y *Mackinnon* para la comparación de dos modelos de regresión que explican la misma variable dependiente. El modelo tendrá que ser escrito de la forma indicada en el manual de usuario del programa.
- El valor numérico *lambda*, comprendido entre 0 y 1, se corresponde con el valor del parámetro definido en el modelo de *Ridge*.
- Por último, al igual que en el resto de funciones, se recibe la variable ruta, que indica la dirección local en la que se encuentran las librerías adicionales usadas por la función, y que no se encuentran por defecto en la instalación típica de R.

El resultado de la función tiene formato lista, en la que todos los elementos de ésta tienen formato matriz, salvo el último que es un vector de elementos de tipo cadena con los nombres de los archivos gráficos generados, o un elemento de tipo *NA*, en caso de no solicitar ninguno.

En primer lugar, la función extrae los valores de las variables de las listas, define correctamente el tipo de cada variable, controla la existencia de multicolinealidad entre las variables cuantitativas seleccionadas, construye las variables *dummies* en el caso de que se hayan incluido variables cualitativas en el modelo, se controla la existencia de multicolinealidad entre ellas, y por último se controla la existencia de multicolinealidad entre la variable dependiente y las variables independientes que se incluirán en el modelo.

A continuación, se seleccionan al azar las muestras de validación y predicción, siempre que el usuario lo haya seleccionado. Una vez hecho esto, se estima ya el modelo de regresión según la especificación indicada, y se generan tres matrices con los resultados de la estimación: la primera de ellas con medidas de bondad de ajuste - coeficiente de determinación y coeficiente de determinación ajustado - y el contraste F de validación global del modelo; la segunda con la estimación de los coeficientes del

modelo, el error estándar de cada uno, así como sus respectivos contrastes t de relevancia; la última de las tres contiene un resumen descriptivo de los residuos del modelo.

Una vez estimado el modelo de regresión los resultados obtenidos, en función de las opciones seleccionadas por el usuario son:

- Medidas de bondad del ajuste. Se calcula el criterio de información de *Akaike*, el criterio de información bayesiano o de *Schwarz* y un criterio de información con un factor de penalización determinado por el usuario.
- Medidas de la capacidad predictiva del modelo. Se calculan, para la muestra de predicción determinada, el error absoluto medio, el error relativo medio y el error cuadrático medio.
- Para el nivel de confianza determinado, se calculan los intervalos de confianza de los coeficientes del modelo, presentándolos en una matriz de tantas filas como coeficientes tenga el modelo estimado, y dos columnas, una para cada extremo del intervalo.
- Gráfico de residuos parciales. Se realiza, para cada variable explicativa del modelo, un gráfico de dispersión en el que se representa en el eje x la variable regresora seleccionada, y los residuos obtenidos al eliminar esta variable del modelo en el eje y. Se representa con la línea de predicción, y sirve para analizar la linealidad entre las variables independientes y la dependiente. Para la obtención del gráfico se ha utilizado la función *prplot* del paquete *faraway*, que ha sido implementado por Faraway (2016) y está desarrollado en su libro. El nombre de estos gráficos será *Parcial.residuos*, seguido del nombre de la variable regresora en estudio, y seguido de un número aleatorio y la extensión del archivo, *jpg*.
- A continuación, se realiza el contraste *RESET*, o test de *Ramsay* para el estudio de la forma funcional del modelo. Para su cálculo, se utiliza la función *RESETTEST* de la librería *lmtest*, desarrollado por Hothorn et al. (2015) y disponible con licencia *GPL* en el repositorio *CRAN* de R. Se devuelve el estadístico de contraste F, los grados de libertad de la distribución, así como su probabilidad límite asociada.
- A partir de la forma funcional alternativa dada por el usuario, se realiza el test de comparación de modelos de *Davidson* y *Mackinonn*. En este, se incluye como variable independiente en uno de los modelos, la variable dependiente estimada por el otro modelo. Si esta variable es relevante, se determina la no idoneidad de la formulación. Se ha vuelto a utilizar para su cálculo una función contenida en el paquete *lmtest* de Hothorn et al. (2015).
- Los tres resultados siguientes están dirigidos al estudio de la existencia de multicolinealidad en el modelo estimado. En el primero de ellos se calculan los índices *VIF* o factores de inflación de la varianza. Se utiliza en esta ocasión la función *vif* del paquete car, creado por Fox et al. (2016) con licencia *GPL* mayor o igual que 2. En los dos siguientes, se muestran el número de condición y el número de condición de la matriz de datos normalizada. En este caso los cálculos se realizan a partir de las funciones básicas de R. Antes de realizar el cálculo de

todos ellos, se realiza el control sobre el número de variables explicativas en el modelo.

- A continuación, se contrasta la hipótesis de normalidad de los residuos mediante tres pruebas posibles: *Shapiro – Wilks*, *Kolmogorv – Smirnov* y *Jarque* – Bera. El segundo de ellos se calcula mediante la aplicación de la función *lillie.test* del paquete *nortest*, de Gross y Ligges (2012), y el tercero se calcula aplicando la función *jarque.bera.test* del paquete *tseries*, cuya autoría es de Trapletti y Hornik (2017). El contraste de *Sharpiro – Wilks* tiene una limitación en el tamaño de la muestra, si ésta es superior a 5000 la función devuelve un error. Esto se ha solucionado mediante la selección, en este caso, de una muestra aleatoria de tamaño 5000, a la que se le aplica el test.
- La normalidad de los residuos también es analizada mediante un *gráfico QQ*, en el que se representan en el plano los puntos dados en el eje x por los cuantiles de la distribución Normal, y en el eje y los dados por la distribución de los residuos. Cuánto más próximos estén los puntos a la bisectriz del primer cuadrante, mayor será la certeza de que la distribución de los residuos se aproxima a una distribución Normal. El nombre de este gráfico será *ResiduosQQ*, seguido de un número aleatorio y la extensión del archivo, *jpg*.
- Para finalizar el análisis de la normalidad de los residuos, se representa un histograma de éstos, que incluye la línea que representa la función de densidad teórica. Los intervalos se han calculado siguiendo la regla de Sturges (1926). El nombre del archivo generado es *ResiduosHist*, seguido de un número aleatorio y su extensión, *jpg*.
- En el siguiente resultado se analiza la existencia de heterocedasticidad en los residuos mediante dos contrastes de hipótesis: el contraste de *Goldfeld* y *Quandt*, y el de *Breusch* y *Pagan*. En ambos se devuelve el estadístico de contraste y la probabilidad límite, todo en la misma matriz de resultados. El contraste de *Goldfeld – Quandt* tiene una limitación en cuanto al ratio entre el número de variables explicativas en el modelo y el tamaño muestral, no se puede calcular para ratios pequeños, ya que éste divide la muestra en tres partes y necesita que el número de datos en cada una de ellas sea suficiente. Este control es realizado durante la ejecución de la función y en caso de no cumplirse la condición para su aplicación es devuelto un mensaje de error en la ventana de resultados. Ambos contrastes se realizan haciendo uso de funciones contenidas en el paquete *lmtest* de Hothorn et al. (2015). Cabe destacar que se ha omitido el contraste de heterocedasticidad de *White* debido a la posible presencia de variables *dummies* en el modelo, que al ser variables que únicamente toman los valores 0 y 1, éstas son invariantes frente a la potenciación, lo que generaría multicolinealidad exacta en el modelo auxiliar de White.
- Este estudio se puede realizar, también, mediante los dos siguientes gráficos que devuelve la función: el gráfico de dispersión entre los residuos del modelo y los valores de la variable dependiente, y el que los compara con cada una de las variables explicativas del modelo. Los nombres con los que se almacenan dichos gráficos son, respectivamente *ResiduosY* seguido del número aleatorio y la extensión *jpg*, y *ResiduosEquis* seguido del índice que índica el orden de la variable explicativa en el modelo, y del número aleatorio y la extensión.
- Los gráficos generados a continuación son similares a los descritos en el punto anterior. La diferencia estriba en considerar los errores cuadráticos, que nos dan

idea de la variabilidad de éstos. Los nombres de los ficheros son similares, *Residuos2Y* y *Residuos2Equis*, continuando la estructura de la misma forma.

- Continuando con el estudio de la validación, y aunque la mayoría de los estudios que se pueden considerar son de tipo transversal, se analiza la existencia de autocorrelación mediante dos pruebas: el contraste de *Durbin – Watson* y el contraste de *Breusch – Godfrey*. Para ambos, vuelve a usarse la librería *lmtest* de Hothorn et al. (2015). Los resultados se devuelven en una única matriz que contiene los estadísticos y las probabilidades límite de ambos contrastes.
- Como último gráfico se propone el gráfico de dispersión que enfrenta los valores de los residuos contra las estimaciones de la variable dependiente. El nombre de este fichero es *PredRes* seguido del número aleatorio y la extensión habitual.
- Para finalizar se proponen dos modelos de regresión que permitan corregir el incumplimiento de alguna de las hipótesis a priori sobre éste: la multicolinealidad y la heterocedasticidad de los residuos. Para el primero de ellos, se propone el modelo de *Ridge* para un valor de λ dado por el usuario. Para el segundo, y haciendo uso de la librería car de Fox et al. (2016) se propone el modelo con corrección de la heterocedasticidad de *White*. No obstante, y como se ha comentado anteriormente, la presencia de variables *dummies* hace inviable el cálculo del nuevo error estándar de los coeficientes.

7.6. Función Regresion2

El archivo que contiene esta función es *regresion2cod.R*. Es llamada por el usuario a través de la pestaña Regresión 2 de la ventana Estudios estadísticos. Para ello, éste deberá, al igual que en la función *regresion*, seleccionar la variable dependiente, entre las variables cuantitativas, y las variables dependientes del modelo, entre las variables cuantitativas y cualitativas.

El objetivo de esta función es analizar, para el conjunto de variables implicadas en un modelo de regresión, la existencia de observaciones anómalas en alguna de las variables o en los residuos del modelo, y observaciones influyentes en la estimación del modelo de regresión.

Las variables cualitativas seleccionadas, de la misma forma que en la función anterior, son introducidas en el modelo de regresión a través de la creación de variables artificiales. La identificación de observaciones atípicas e influyentes se realiza a través de la variable Código, que es única para cada inmueble.

Los argumentos de la función son:

- El vector lógico *opcionesL*, de longitud 27, que recoge las opciones seleccionadas por el usuario a través de la ventana.
- La tabla de datos *depend*, que a diferencia de la función *regresion*, está compuesta por dos columnas, la primera de las cuales se corresponde con la variable código, y permitirá identificar las observaciones. La segunda variable incluida es la variable dependiente del modelo de regresión. Como en el resto de ocasiones, es recibida por JAVA en formato lista, de la que hay que extraer la tabla de datos.
- La tabla de datos *independV* que incluye las variables cuantitativas seleccionadas por el usuario para formar parte del conjunto de variables regresoras del modelo.
- La tabla de datos *independF*, que contiene el conjunto de variables cualitativas a partir del cual se construirán las variables artificiales que se incluirán en el modelo.
- La variable de tipo cadena *form*, que contiene la formulación del modelo de regresión, distinta a la natural, en caso de que esta opción sea seleccionada.
- La variable de tipo real denominada muestra, que contiene el tamaño en términos relativos o absolutos de la muestra de validación.
- La variable *limtip* determina el número de veces la desviación típica que tiene que alejarse el valor de una variable de su media para ser considerada una observación atípica u *outlier*.
- De la misma forma se define la variable *limrst*, para el caso en el que se estudia la existencia de valores atípicos en los residuos *estudentizados*.
- La variable *limhat* es propuesta por el usuario para determinar el grado de apalancamiento mínimo de una observación para que ésta sea considerada una observación influyente. El usuario puede elegir el límite dado por defecto por el programa. En ese caso:

$$limhat = 2 \cdot \frac{k+1}{n}$$

  Donde k es el número de variables regresoras y n el tamaño muestral.

- El siguiente argumento de la función es *limdfb*, que es la distancia mínima, indicada por el usuario, entre el valor estimado de alguno de los coeficientes del modelo de regresión con todas las observaciones y el obtenido al eliminar una determinada observación, para considerar ésta influyente. El valor por defecto que se utiliza en caso de no ser facilitada es:

$$limdfb = \left|\frac{2}{\sqrt{n}}\right|$$

- La variable *limdff* determina el valor – frontera para la distancia entre el valor estimado para una determinada observación teniendo en cuenta todas las observaciones, y el estimado al eliminar dicha observación. Si este valor es superado, la observación es considerada influyente. El valor utilizado por defecto es:

$$limdff = \left| 2 \cdot \sqrt{\frac{k+1}{n}} \right|$$

- Otra medida de influencia es la distancia de *Cook*, y el límite impuesto por el usuario se guarda en la variable *limdc*. Esta es la distancia mínima que debe haber entre el vector de predicción de todas las observaciones de estudio utilizando para ello todas las observaciones, y el obtenido al eliminar una observación concreta. En ausencia de un valor indicado por el usuario, éste es:

$$limdc = \frac{4}{n}$$

- La variable *limcov* indica el valor por encima o por debajo de 1 que puede tomar el *COVratio*, que es el cociente entre los determinantes de las matrices de covarianzas de los coeficientes al incluir todas las observaciones y eliminar una concreta, para considerar una observación influyente. El valor considerado en ausencia de otro indicado por el usuario es:

$$limcov = \frac{3 \cdot (k+1)}{n}$$

- El valor numérico *limAIC* es el número máximo de desviaciones típicas que se puede desviar el valor del criterio de información obtenido al eliminar una determinada observación respecto al valor medio de todos los valores del criterio de información resultantes al eliminar una observación cada vez. Este valor es enviado tanto para el método de detección de observaciones como para el método de selección, y también para el criterio de información de *Akaike*, el de *Schwarz* y cualquier otro factor de penalización. Este factor es enviado a través del argumento pen.
- La variable numérica *alpha* es un valor comprendido entre 0 y 1, que se corresponde con el parámetro del modelo de *Ridge*, es decir:

$$\hat{\beta}_{RIDGE} = (X^t \cdot X + \lambda \cdot I)^{-1} \cdot X^t \cdot Y$$

- Por último, y como en todas las funciones, se envía la ruta de acceso local a las librerías de R necesarias para la correcta ejecución de la función. La variable es *ruta*.

El resultado de la función regresion2, es una lista de elementos de tipo matriz, salvo el último de los elementos que es el vector de cadenas, que incluye los nombres de los archivos creados.

La preparación de los datos por parte de la función es similar al realizado en la función *regresion*: extraer de las listas las tablas de datos, redefinirlas según el tipo, crear las

variables artificiales a partir de las variables cualitativas, y controlar la presencia de multicolinealidad exacta.

Una vez hecho esto los resultados de la función son los siguientes:

- En primer lugar, para la variable dependiente, y cada una de las variables cuantitativas regresoras, se analiza la existencia de valores atípicos mediante dos métodos: el método de los cuartiles, en el que se consideran anómalo todo aquél que esté por encima del tercer cuartil o por debajo del primer cuartil, más de tres veces el rango intercuartílico; y el método de tipificación, por lo que se considera valor atípico todo aquél que se desvíe de la media más de *limtip* veces la desviación típica.
- A continuación, se estudia la existencia de valores atípicos en los residuos del modelo. Para ello, se realiza el test de valores atípicos de *Bonferroni* y se analiza la existencia de residuos *estudentizados* que se desvíen un número de veces la desviación típica de la media, impuesto por el usuario y que se almacena en la variable *limrst*. Para el cálculo del test de *Bonferroni* en la detección de observaciones atípicas se ha utilizado el paquete car de Fox et al. (2016), en concreto, la función *outlierTest*.
- Un *gráfico QQ* para los residuos del modelo que nos permite analizar gráficamente la existencia de valores atípicos. El nombre generado es *QQresid* acompañado de un número aleatorio de hasta 4 cifras y la extensión del archivo.
- Matriz de datos que contiene, en filas, las observaciones con efecto palanca mayor que el determinado por el usuario a través de la variable *limhat*. La primera columna de la matriz muestra el valor palanca obtenido, y el resto los valores de la variable dependiente e independientes cuantitativas. Los nombres de cada fila se corresponden con los códigos de los inmuebles mostrados.
- El estudio anterior se realiza gráficamente a continuación, mediante los gráficos *Leverages*, en los que se representan los gráficos de dispersión entre la variable dependiente y cada una de las variables independientes. Se muestran todos los gráficos en un solo archivo siempre que el número de variables independientes sea inferior a 10, debido a limitaciones del software R. En caso contrario, se representan varios gráficos. El nombre del archivo o archivos es *Leverages*, seguido de un número aleatorio de a lo sumo cuatro cifras, y su extensión. Para esta representación gráfica se ha utilizado de nuevo la librería car de Fox et al. (2016).
- Matriz de datos que contiene las observaciones con valores *DFFITS* superiores al límite establecido por *limdff*, es decir, aquellas en las que la diferencia entre la estimación de su valor a través del modelo difiere significativamente de la obtenida excluyendo dicha observación de la muestra. El formato de matriz devuelto es igual al anterior.
- La siguiente matriz contiene información de las observaciones en las que al menos el valor *dfbeta* tipificado de uno de los coeficientes del modelo es mayor que el especificado por *limdfb*. Se muestran los valores *dfbeta* normalizados de todos los coeficientes del modelo de las observaciones seleccionadas. Los

nombres de cada fila están formados por los códigos de los inmuebles resultantes del análisis.

- Matriz que contiene las observaciones cuya distancia de *Cook* es superior a la máxima dada por *limdc*. La estructura es similar a la devuelta en el estudio de los valores *dffits* o los valores con efecto palanca: Los nombres de las filas se corresponden con los códigos de los inmuebles, la primera columna muestra los valores de la distancia de *Cook* de cada uno de ellos, y en el resto de columnas se muestran los valores de la variable dependiente y de las independientes cuantitativas de dichas observaciones.
- El siguiente resultado es un gráfico en el que se muestran los valores de la distancia de *Cook* de todas las observaciones . En éste, además, se muestra la línea frontera que delimita el valor de la distancia de *Cook* a partir de la cual se considerarán observaciones influyentes. El nombre de este archivo es *CookD*, seguido de un número aleatorio y la extensión del archivo.
- El siguiente método para detectar observaciones influyentes, es el cálculo del *COVratio* de cada observación, que es el cociente entre el determinante de la matriz de covarianzas de los coeficientes y el de la misma matriz al eliminar la observación que se está analizando. Si este cociente es 1, la observación no es influyente. La matriz - resultado está compuesta por las observaciones cuyo *COVratio* se aleja de 1 una cantidad mayor que *limcov*. Se indica también el código del inmueble, el valor del *COVratio*, y los valores de la variable dependiente y las independientes cuantitativas.
- A continuación, se representan tres gráficos con el nombre *influenciALL*, en los que se resume el estudio realizado hasta este momento. Los gráficos mostrados son: el gráfico de los valores palanca, el gráfico de residuos *estudentizados* y el gráfico de valores de probabilidad límite del test de *Bonferroni*. De nuevo es necesario el uso de la librería car de Fox et al. (2016).
- El siguiente resultado es el que utiliza la técnica propuesta en este trabajo para la detección de observaciones influyentes, que consiste en comparar los valores del criterio de información del modelo estimado al eliminar una observación del conjunto de datos. Se ofrece la matriz que contiene las observaciones consideradas influyentes a partir de este método, siendo éstas aquellas para las que el valor del criterio de información al eliminar dicha observación se aleja por defecto de la media más de *limAIC* veces la desviación típica hallada. Esta matriz tiene la misma composición que las anteriormente definidas, mostrando el código y los valores de las variables para dicha observación. El factor de penalización usado para el cálculo del criterio de información es seleccionado previamente por el usuario.
- Se incluye además un gráfico en el que se muestran los valores obtenidos para el criterio de información cuando se elimina cada vez una observación del conjunto de datos. El archivo gráfico se denomina *AIC*, seguido de un número aleatorio y la extensión
- Las últimas siete matrices resultado muestran la información de la estimación del modelo original, y el resultante al aplicar el algoritmo propuesto de selección de observaciones, a partir del valor del criterio de información. En este se calculan los n valores de *AIC* de los modelos estimados al eliminar cada vez una

observación, y se halla el valor medio y la desviación típica de éstos. Si todos los valores son superiores a:

$$\overline{AIC} - limAIC \cdot s_{AIC}$$

Finaliza el proceso. En caso contrario se elimina del conjunto de datos la observación correspondiente al menor valor de AIC, y se repite el proceso calculando los $n - 1$ valores de AIC resultantes al eliminar una observación.

La primera de las matrices resultantes contiene: el coeficiente de determinación, el coeficiente de determinación ajustado y el contraste de validación global del modelo original. La segunda matriz contiene los coeficientes del modelo original estimado junto con sus correspondientes contrastes de relevancia. La última matriz contiene un resumen descriptivo de los residuos de este modelo. La siguiente matriz muestra las observaciones eliminadas durante el proceso de selección. Las tres últimas replican las tres primeras para el modelo resultante de eliminar dichas observaciones.

7.7. Función Perceptron

El código de esta función está contenido en el archivo *perceptron.R*. Se ejecuta a través de la pestaña Redes Neuronales de la ventana Estudios estadísticos. De la misma forma que en las funciones anteriores *regresion* y regresion2, el usuario deberá seleccionar qué variable cuantitativa será la que se desea estimar y por tanto la variable dependiente del modelo, así como el conjunto de variables que se utilizarán en esta estimación, es decir, el conjunto de variables independientes del modelo, que serán elegidas entre el conjunto de variables cuantitativas restantes y / o el conjunto de variables cualitativas.

En esta ocasión, la estimación se realizará a través de la metodología de redes neuronales, concretamente a través de la aplicación de un perceptrón multicapa, cuya arquitectura será seleccionada por el usuario.

Al igual que como se ha explicado en funciones anteriores, las tablas de datos son extraídas de las listas en las que se reciben, se construyen las variables artificiales a partir de las variables cualitativas seleccionadas y se controla la no existencia de multicolinealidad exacta ni de relación perfecta entre la variable dependiente y alguna de las independientes.

Los argumentos que necesita la función *Perceptron* para su ejecución son los siguientes:

- Las tablas de datos en formato lista denominadas y, *equisV* y *equisF*; que contienen respectivamente los datos de la variable dependiente, las variables cuantitativas y las cualitativas seleccionadas.
- Las tablas de datos, también en formato lista, correspondientes a nuevos valores de las variables independientes - cualitativas y cuantitativas - para los que se desea estimar el valor de la variable dependiente. La función controla que los valores recibidos por cada variable cualitativa estén contenidos en los valores de ésta, en las observaciones de entrenamiento de la red. Se denominan respectivamente *equisPF* y *equisPV*, y llegan también incluidos en una lista, por lo que se extrae de ésta con anterioridad a la estimación de la red.
- El vector lógico *opcL* es un vector de longitud 12, que contiene la información aportada por el usuario a la función a través de sus preferencias.
- La muestra se divide en dos: la denominada muestra de entrenamiento de la red, y la muestra de validación. El tamaño de ésta última se transmite a la función por el usuario a través de la variable *muestra*, de tipo numérico.
- La variable *form*, de tipo cadena, contiene la formulación de un modelo alternativo al modelo lineal básico generado con las variables seleccionadas.
- El vector *neurocapas* contiene la información básica de la arquitectura de la red. Es un vector con una longitud comprendida entre 1 y 4, que coincide con el número de capas ocultas de la red. Cada componente de dicho vector contiene el número de neuronas de cada una de las capas definidas.
- La variable *grer* es el valor numérico que especifica el umbral de las derivadas parciales de la función de error establecido como criterio de parada. Es decir, la variación de error máxima entre dos iteraciones consecutivas para detener las iteraciones del método.
- Otro criterio de parada que debe imponerse en el entrenamiento de la parada para evitar que éste diverja es el número máximo de iteraciones a realizar. Este valor numérico se guarda en la variable *max.it*. El procedimiento se detendrá siempre que se dé uno de los dos criterios de parada: Se alcanza el umbral de la derivada parcial del error o se llega al número máximo de iteraciones.
- En el caso de que se seleccione una única capa oculta en la red, se pueden calcular los intervalos de confianza de los pesos de la red óptima. En este caso, el nivel de confianza con el que se calculan estos intervalos se almacena en la variable *nivel*.
- Por último, la variable *ruta* contiene la dirección local en la que se encuentran las librerías adicionales necesarias para la ejecución de la función, y sus correspondientes dependencias.

El resultado de la función es una lista de matrices que contienen los resultados de la estimación de la red, junto a un conjunto de gráficos que expondremos a continuación, y un último elemento que es un vector de elementos tipo cadena con los nombres de los ficheros de gráficos generados.

La asignación de pesos iniciales de la red es aleatoria, por lo que dos ejecuciones consecutivas de esta función no devolverán resultados idénticos. Es más, es posible que una ejecución sea convergente mientras que la otra no lo sea. En este último caso, el resultado es una lista con sólo dos vectores. En el primero de ellos se indica que se ha llegado al número máximo de iteraciones sin que el proceso haya convergido, el segundo es un vector con un único valor perdido, que informa al programa que no se han generado archivos gráficos.

Los resultados devueltos por la función en el caso de que la red sea convergente y el usuario active todas las opciones son los siguientes:

- El primer elemento del resultado de la función es una matriz - resumen del resultado obtenido para la red. Este resultado se muestra siempre que ésta converja y sin que el usuario especifique ninguna opción concreta. En esta matriz se muestra el error alcanzado, el umbral, el número de pasos necesarios hasta llegar a la convergencia, los valores de los criterios de información de *Akaike* y *Schwarz* y todos los pesos de la red óptima. Para la construcción de esta matriz se ha hecho de la función *neuralnet* del paquete del mismo nombre, obra de Günther y Fritsch (2010) con licencia *GPL* ≥ 2, y disponible en el repositorio *CRAN* de R.
- Una matriz que incluye el error cuadrático medio de la red y del modelo de regresión equivalente, para la muestra de validación antes seleccionada. Estas medidas de la capacidad predictiva nos permiten comparar ambos métodos.
- A continuación, se devuelve la matriz que incluye los valores de predicción de la variable dependiente para las observaciones solicitadas por el usuario a partir de la red construida y del modelo de regresión equivalente. Estas observaciones se han introducido por el usuario a través de la matriz habilitada para ello en la pestaña Estimación de los valores, contenida en la pestaña Redes Neuronales del programa. Cada fila de la matriz está formada por los valores de predicción y los de las variables independientes de cada observación.
- Las siguientes matrices están compuestas por los pesos de la red óptima. Se obtendrán tantas matrices como el número de capas ocultas de la red más 1, ya que en cada una de ellas se muestran los pesos entre cada neurona de la capa k y las de la capa k+1.
- Si la red se compone de una única capa oculta, el usuario puede seleccionar la opción de calcular los intervalos de confianza de los pesos de ésta con un nivel de confianza también a su elección. El resultado se muestra en cinco matrices, las dos primeras para los extremos inferiores de los intervalos, las dos últimas para los extremos superiores, y la última es una matriz compuesta por un único elemento que representa el denominado criterio de información de la red (NIC). En total se calculan $k \cdot m + m \cdot 1$ intervalos de confianza, donde k es el número de variables explicativas, y m es el número de neuronas de la única capa oculta definida. En ocasiones, fundamentalmente cuando el tamaño muestral es reducido, no es posible calcular estos intervalos debido a la existencia de neuronas irrelevantes. Para el cálculo de éstos intervalos se ha hecho uso de la función

confidence.interval, también perteneciente al paquete *neuralnet* de Günther y Fritsch (2010).

- A continuación, se representa gráficamente la arquitectura de la red indicando el valor de cada peso en la red óptima. Éste gráfico se almacena con el nombre de archivo red, seguido de un número aleatorio y de la extensión *jpg*.
- Los siguientes gráficos representan los pesos generalizados de la red para cada una de las variables explicativas del modelo, tomando como variable respuesta la dependiente. Por tanto, se representa tantos gráficos de pesos generalizados como número de variables explicativas se hayan seleccionado. El nombre de los gráficos es *pesosvar* seguido de un índice que indica la variable explicativa para la que se ha realizado el gráfico, un número aleatorio y la extensión del archivo.

7.8. Función Mapeado

El archivo que contiene el código fuente de esta función se denomina mapa.R. Se ejecuta desde la pestaña Mapa de la ventana de Estudios estadísticos del programa. Esta función tiene como objetivo mostrar, en una imagen satélite, la situación geográfica de los inmuebles de la muestra, a partir de la longitud y la latitud de cada inmueble. Además de la localización de los inmuebles en el espacio, la diferencia de color de los puntos que los representan, indicará el valor de la variable analizada.

En la elección de la variable se distingue entre el grupo de variables cualitativas, en el que se elegirá exactamente una de ellas, entre las disponibles, que son: *Dirección, Barrio, Distrito, Municipio, Provincia, Tipo, Piso, Cochera, Ascensor, Piscina, AireAcon, Terraza, Trastero, Tendederos, Empotrados, Estado, Localización, Esquina, Salida de Humos, Seguridad, Distribución, Uso del Edificio, Agua Caliente Independiente, Calefacción independiente, Puerta Automática y Parking de Motocicletas*, en función del tipo de inmueble estudiado.

En el grupo de variables cuantitativas puede seleccionarse una variable para representar, o dos variables, en el caso en el que se desee representar una operación entre ellas, por ejemplo, el precio de un inmueble por unidad de superficie. Las variables disponibles son: *Longitud, Latitud, Distancia, NumHab, NumBaños, Tamaño, Precio, NumFotos* y las correspondientes a las distintas actualizaciones de los precios de un inmueble.

Los argumentos de la función son los siguientes:

- Una tabla de datos, que es recibida en formato lista, que contiene en las dos primeras columnas, los valores de longitud y latitud de los inmuebles, y que contiene, además, una o dos columnas con los valores a representar a través de los códigos de color.
- Un vector lógico denominado *opcL*, que contiene las preferencias seleccionadas por el usuario en la elaboración del mapa.
- El valor numérico *Mayor*, que contiene el máximo valor de la variable que se desea representar, en el caso en el que se limite superiormente el valor de ésta.
- El valor numérico *Menor*, que contiene el mínimo valor de la variable que se desea representar, en el caso en el que se limite inferiormente el valor de ésta.
- El valor numérico *zum*, que representa el tamaño del mapa, y, por tanto, la cantidad de superficie recogida por éste. Aunque el valor de este no se encuentra limitado por la función, sólo puede tomar valores enteros entre 10 y 18. No obstante la aplicación la realiza este control.
- El valor numérico *alfa* determina la opacidad de los puntos en el mapa. Su valor se restringe en la aplicación a valores entre 0 y 1, donde el valor 0 se corresponde a la transparencia total, y el 1 a la opacidad completa.
- El valor numérico *tamp* determina el tamaño de los puntos en el mapa. Desde la aplicación se restringe su tamaño a valores entre 1 y 10. Este tamaño es independiente del zoom del mapa.
- La variable de tipo cadena con la ruta en la que se encuentran los paquetes adicionales necesarios para la ejecución de la función.

El resultado de la función es un elemento de tipo cadena que indica el nombre del archivo *jpeg* generado con el mapa. La descarga del mapa desde el repositorio de Google y la generación de los puntos en éste, se realiza gracias al paquete ggmap, desarrollado por Kahle y Wickham (2013) y un gran número de paquetes de los que tiene dependencias entre los que destaca el paquete ggplot2 de Wickham (2009).

El uso de un gran número de paquetes o librerías ya implementadas en R, han hecho posible la creación de las funciones previamente explicadas. Además de las librerías integradas en la instalación básica del programa, se ha hecho uso, directamente o a través de dependencias, de las librerías *car, colorspace, digest, e1071, faraway, geosphere, ggmap, ggplot2, gtable, images, jpeg, kendall, labeling, lme4, lmtest, magrittr, mapproj, maps, MASS, MatrixModels, minqa, munsell, neuralnet, nloptr, nortest, pbkrtest, plyr, png, ppcor, proto, quadprog, quantreg, Rcpp, reshape2, RgoogleMaps, rjson, RJSONIO, scales, sp, SparseM, stringi, stringr, tseries y zoo*.

Todas ellas se incluyen en la instalación del programa.

8.Análisis del mercado inmobiliario de Sevilla

En el siguiente capítulo se realizan dos estudios para el mercado inmobiliario de oferta de la provincia de Sevilla a partir de la base de datos contenida en el portal inmobiliario Idealista, recogida y analizada con el programa propuesto en este trabajo.

8.1. Análisis descriptivo del territorio

Tras un detalle de situación de la provincia en la que se ha realizado el estudio, se realiza un primer estudio del mercado de viviendas a la venta en el municipio de Sevilla. A continuación, se analizan los locales comerciales a la venta contenidos en aglomeración urbana de Sevilla, denominado así al conjunto formado por el municipio de Sevilla, y los municipios Albaida del Aljarafe, Alcalá de Guadaíra, Alcalá del Río, La Algaba, Almensilla, Aznalcázar, Aznalcóllar, Benacazón, Bollullos de la Mitación, Bormujos, Brenes, Camas, Carmona, Carrión de los Céspedes, Castilleja de Guzmán, Castilleja de la Cuesta, Castilleja del Campo, Coria del Río, Dos Hermanas, Espartinas, Gelves, Gerena, Gines, Guillena, Huévar del Aljarafe, Isla Mayor, Mairena del Alcor, Mairena del Aljarafe, Olivares, Los Palacios y Villafranca, Palomares del Río, Pilas, La Puebla del Río, La Rinconada, Salteras, San Juan de Aznalfarache, Sanlúcar La Mayor,

Santiponce, Sevilla, Tomares, Umbrete, Utrera, Valencina de la Concepción, Villamanrique de la Condesa, Villanueva del Ariscal y El Viso del Alcor.

8.1.1. La provincia de Sevilla

En el centro de la comunidad autónoma de Andalucía se encuentra la provincia de Sevilla. Limita con las provincias de Córdoba, Badajoz, Huelva, Cádiz y Málaga. Tiene una extensión de 14.036 kilómetros cuadrados, repartida en 105 municipios. Su capital es la ciudad de Sevilla, con una extensión de 141,3 km^2. El municipio con mayor extensión es Écija, con 978,7 km^2 y el de menor, Castilleja de Guzmán, con tan sólo 2,1 km^2. El más próximo a la capital es Camas, a tan sólo 5 km de distancia, y el que más alejado se encuentra es Badolatosa, a 135 kilómetros de ésta. (Diputación de Sevilla, 2016). Un mapa con la distribución de los municipios existentes en la provincia puede observarse en la Figura 125.

Figura 125. Mapa de la provincia de Sevilla. (Fuente: Diputación de Sevilla)

Según la última información publicada en el instituto nacional de estadística (INE), en 2016, la población total de la provincia, asciende a 1.939.775 personas, con un ligero predominio de mujeres respecto a hombres. Esta población representa algo más del 23 % de la población total de la comunidad autónoma de Andalucía, y su crecimiento actual se encuentra en valores negativos.

En cuanto a la densidad de población, la provincia, con 138,32 habitantes por kilómetro cuadrado, se sitúa en niveles superiores a los registrados a nivel nacional y autonómico, en valores por debajo de los 100 habitantes por kilómetro cuadrado. Esta información puede consultarse en el anuario estadístico de la diputación de Sevilla, Diputación Provincial de Sevilla (2016)

Respecto a la economía de la provincia, y según los últimos datos publicados por el INE (2016) - correspondientes al año 2014, el producto interior bruto experimentó un incremento del 1,83 %, respecto al año anterior, situándose en algo más de 34.873 millones de euros. Las actividades principales fueron las relacionadas con el sector público, seguidas del comercio al por mayor y al por menor; reparación de vehículos de motor y motocicletas; transporte y almacenamiento; hostelería; información y comunicaciones.

Hay, en la provincia un total de 113.653 empresas, de las cuales, el 43,4 % pertenecen al sector servicios, y un 41,6 % al comercio, transporte y la hostelería. El número de empresas del sector servicios dedicadas a actividades inmobiliarias es de 4.697, lo que supone un 9,52 % de este sector, y un 4,13 % del total.

La tasa de actividad es del 57,92 % y la de paro, del 28,02 %, según la encuesta de población activa. En total, el número de ocupados de la provincia asciende a 658.100 trabajadores.

Una vez analizadas las principales características de la provincia de Sevilla, profundicemos en las características del mercado inmobiliario. Los datos expuestos a continuación han sido extraídos de la base de datos ministerial, Fomento (2016).

Se estima que en 2016 existía un total de 897.645 viviendas en la provincia de Sevilla, y este valor ha ido creciendo anualmente desde el año 2001, como puede apreciarse en la

Figura 126, aunque este incremento se ha amortiguado en los últimos años con motivo de la crisis inmobiliaria.

El número de viviendas libres iniciadas durante el año, 2015, fue de 1.316, muy lejos del récord histórico de la serie que data del año 2006 y que ascendió a 24.207. Valores similares se obtienen para el número de viviendas libres terminadas, 1.178 en 2015 frente a las 19.498 de 2006.

El precio medio por metro cuadrado de las viviendas libres se situó, en el tercer trimestre de 2016, en 1.245,30 €, continuando la tendencia a la baja, experimentada en los últimos años. No en vano, este precio, se redujo un 1,1 % respecto al mismo trimestre del año anterior.

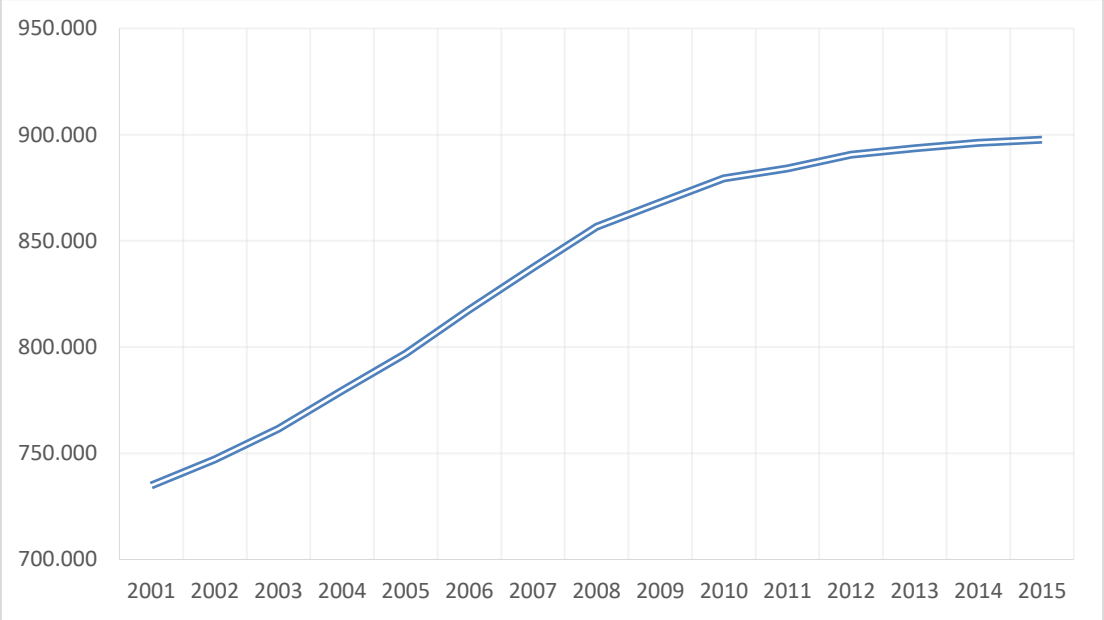


Figura 126. Parque de viviendas de la provincia de Sevilla. (Fuente: Ministerio de Fomento)

La evolución del precio de la vivienda libre de la provincia de Sevilla, se muestra en la Figura 127. Alcanzó su valor máximo en el primer trimestre de 2008, con un valor de 1.791,20 € por metro cuadrado. Desde principio de siglo hasta ese momento, el incremento experimentado por el precio fue elevado. Desde el valor máximo, éste ha ido decreciendo, de forma que el precio actual es un 30 % inferior al máximo registrado.

Las viviendas con una antigüedad igual o inferior a 5 años, el precio es ligeramente superior, situándose en el tercer trimestre de 2016 en 1.426 €, un 3 % inferior al precio

un año antes. Por el contrario, las viviendas de antigüedad superior a los 5 años tienen un precio medio de 1.241,30 €, valor muy cercano al precio medio total, por el mayor número de éstas.

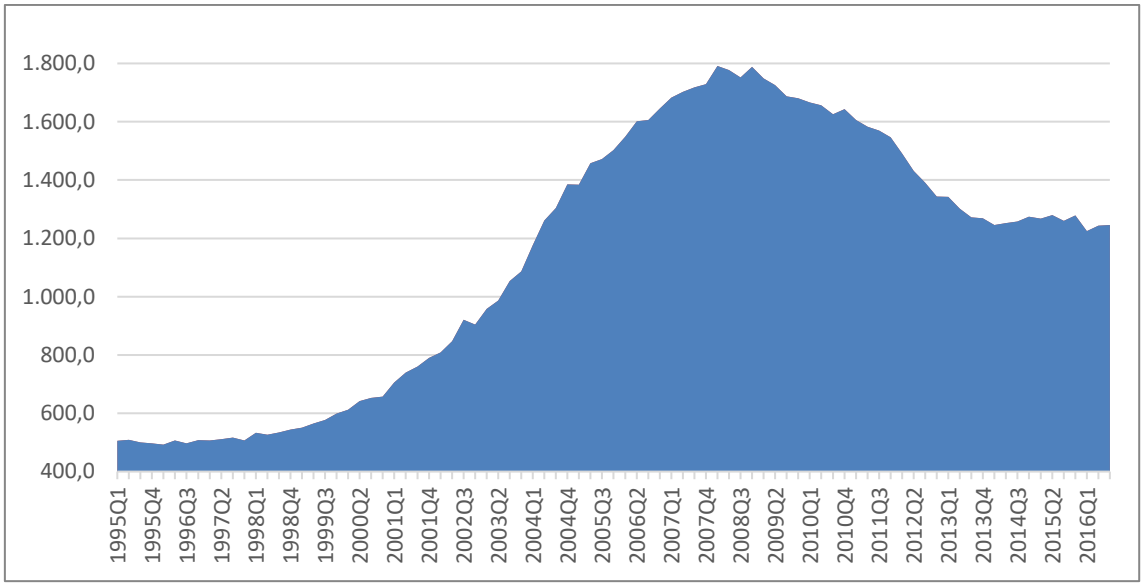


Figura 127. Evolución del precio medio de la vivienda libre en la provincia de Sevilla. (Fuente: Ministerio de Fomento)

En el caso de viviendas protegidas, este precio es de 1.011,20 € por metro cuadrado, con variaciones inferiores en el tiempo que las experimentadas por el precio de la vivienda libre.

Por otro lado, el precio por metro cuadrado del suelo urbano de la provincia tiene un comportamiento similar al de las viviendas, con valores alejados de los máximos históricos. Su valor en el tercer trimestre de 2016 es de 142,7 € por metro cuadrado, mientras que el valor máximo se obtuvo en el cuarto trimestre de 2007 con más de 255 €. No obstante, en el último año, este valor ha aumentado con fuerza respecto al del año anterior, con una subida cercana al 15 %. El valor total de las transacciones de suelo ascendió, en el tercer trimestre de 2016, a 16.869.100 €.

Durante el tercer trimestre del año 2016 se realizaron en la provincia 3.170 transacciones inmobiliarias de viviendas, con un total de 138.700 metros cuadrados, lo que representa un incremento significativo respecto al dato obtenido en el tercer trimestre de 2015, el que se registraron 2.883 transacciones. Este comportamiento se mantiene en ligero crecimiento desde el año 2012, aunque todavía se encuentra alejado del número de

transacciones, de este tipo, que se realizaban en los años 2006 y 2007, momento en el que el número de transacciones se situaba por encima de las 8.000 trimestrales. La evolución del número de transacciones desde el primer trimestre de 2004 hasta el tercer trimestre de 2016, puede observarse en la Figura 128.

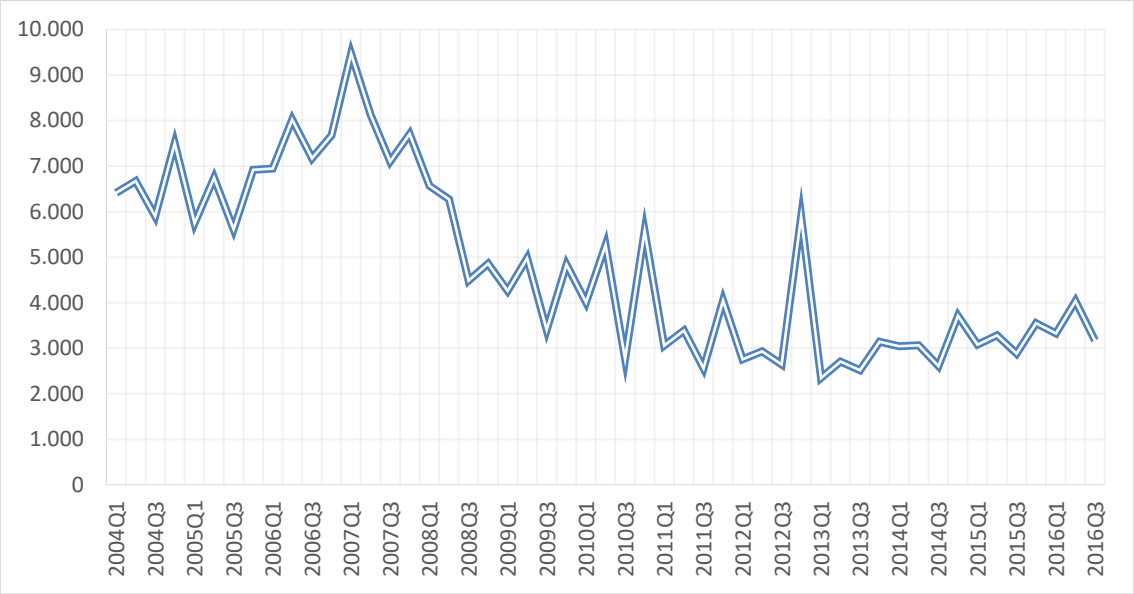


Figura 128. Evolución trimestral del número de transacciones de viviendas en la provincia de Sevilla. (Fuente: Ministerio de Fomento)

Como puede apreciarse, la serie presenta una componente estacional, de forma que el número de transacciones aumenta en el cuarto trimestre, mientras que es menor durante el tercer trimestre del año. También puede observarse el importante incremento producido en diciembre de 2012, motivado por la reforma en la que se eliminó la deducción por vivienda habitual para las viviendas compradas a partir de enero de 2013.

De las 3.170 transacciones de viviendas realizadas en el tercer trimestre de 2016, el 86,5% se corresponden a viviendas libres, y el resto a viviendas protegidas. De las primeras, el 36,8 % se corresponden con viviendas unifamiliares y sólo el 6,34 % a viviendas nuevas, de las que el 61 % son unifamiliares. De las segundas, el porcentaje de viviendas nuevas aumenta hasta el 25,8 %.

El valor medio de las transacciones inmobiliarias de vivienda libre asciende en este mismo trimestre a 108.628 €, con un total de 297.966.500 €, de los que sólo 20 millones se corresponden a vivienda libre nueva, a un precio medio de 114.993 €. Por el contrario,

las transacciones de viviendas libres de segunda mano se realizaron con un valor medio de 108.197€.

La Tabla 25, muestra la distribución de las transacciones de vivienda libre de la provincia de Sevilla en función de la superficie construida y el valor en euros, en el segundo trimestre de 2016, publicada por Fomento (2016).

Tabla 25.Distribución del número de transacciones de viviendas en la provincia de Sevilla según superficie y valor en el segundo trimestre de 2016

	Valor (miles €)								
Superficie (m²)	**[0-150]**	**(150 -300]**	**300-450**	**450-600**	**600-750**	**750-900**	**900-1050**	**1050 o más**	**Total**
Total	2.848	532	86	30	8	4	5	3	**3.516**
menos de 30	44	2	-	-	-	-	-	-	**46**
de 30,01 a 60	486	4	1	1	-	-	-	-	**492**
de 60,01 a 90	1.130	76	1	-	-	-	-	-	**1.207**
de 90,01 a 120	619	171	10	3	1	-	1	-	**805**
de 120,01 a 150	341	112	19	-	-	-	-	-	**472**
de 150,01 a 180	125	64	16	12	-	-	-	-	**217**
más de 180	103	103	39	14	7	4	4	3	**277**

Fuente: Ministerio de Fomento

La mayoría de las transacciones se realizaron sobre viviendas con una superficie superior a los 60 metros cuadrados y hasta los 90 metros cuadrados, con un valor máximo de 150.000 €. Más del 96 % de las transacciones, se realizaron por debajo de los 300.000 €.

8.1.2. La aglomeración urbana de Sevilla

Como se ha comentado anteriormente, la provincia de Sevilla se divide en 105 municipios, en los que reside una población cercana a los 2 millones de habitantes. Tres cuartas partes de esa población reside en la denominada Aglomeración urbana de Sevilla, que reúne los municipios de Albaida del Aljarafe, Alcalá de Guadaíra, Alcalá del Río, La Algaba, Almensilla, Aznalcázar, Aznalcóllar, Benacazón, Bollullos de la Mitación, Bormujos, Brenes, Camas, Carmona, Carrión de los Céspedes, Castilleja de Guzmán, Castilleja de la Cuesta, Castilleja del Campo, Coria del Río, Dos Hermanas, Espartinas, Gelves, Gerena, Gines, Guillena, Huévar del Aljarafe, Isla Mayor, Mairena del Alcor,

Mairena del Aljarafe, Olivares, Los Palacios y Villafranca, Palomares del Río, Pilas, La Puebla del Río, La Rinconada, Salteras, San Juan de Aznalfarache, Sanlúcar La Mayor, Santiponce, Sevilla, Tomares, Umbrete, Utrera, Valencina de la Concepción, Villamanrique de la Condesa, Villanueva del Ariscal y El Viso del Alcor; y que fue aprobado por Decreto 267/2009, de 9 de Junio (BOJA núm. 132, de 9 de julio de 2009).

La Figura 129, muestra una infografía de la Aglomeración urbana de Sevilla, en la que se muestran los principales municipios que la componen. Sobre sombreado gris se indica el núcleo urbano perteneciente al municipio de Sevilla.

Según los datos del Instituto de Estadística y Cartografía de Andalucía, a 1 de enero de 2015, el número de establecimientos en la provincia de Sevilla era de 128.389, mientras que en 2007 este número se situaba en 130.439, lo que supone un descenso de más de un 1,5 %.

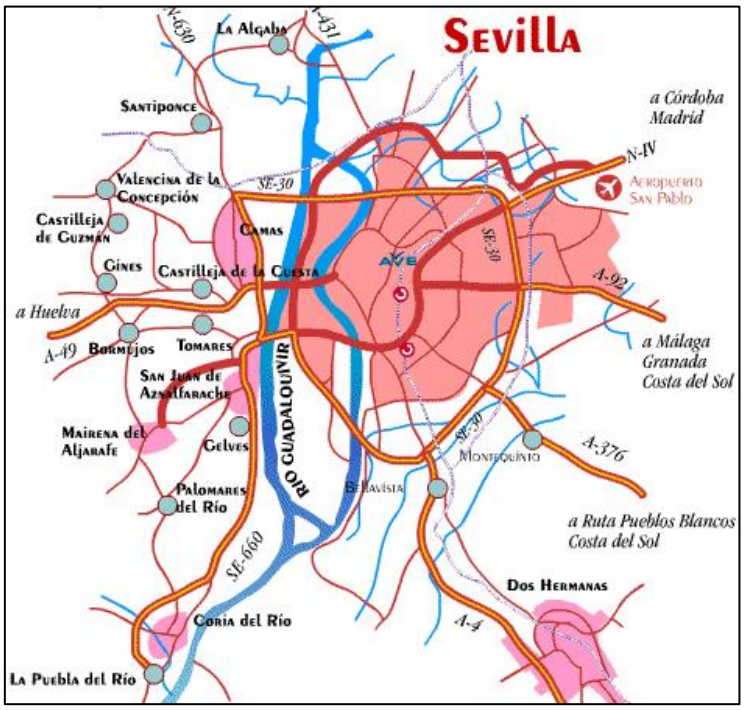


Figura 129. Infografía de la aglomeración de Sevilla. (Fuente: Sevilla.org)

Según el informe de 2015 publicado por *Inerzia*, el precio de las naves industriales del área metropolitana de Sevilla ha sufrido un descenso del 69,1 % desde el año 2007, pasando de un precio medio máximo de 1.046 € por metro a los 506 € del año 2015.

En la zona urbana, los polígonos industriales con mayor precio medio, según este mismo informe, son los de Pagusa, Carretera Amarilla y Store, situados en los distritos

de San Pablo y Santa Clara, con valores por encima de los 730 € por metro cuadrado. No obstante, este precio está muy alejado del encontrado en el polígono industrial Manchón, situado en el municipio de Tomares, con un precio medio de 992 € por metro cuadrado. En el lado contrario se sitúan los polígonos de la Chaparrilla e Hytasa, en Cerro Amate, con valores inferiores a los 450 € por metro cuadrado.

En los municipios colindantes, el precio medio obtenido en este informe se sitúa por debajo del calculado en la zona urbana, salvo en la zona Oeste, donde se sitúa el polígono Manchón antes comentado. Esta zona es la que ha sufrido, además, un mayor descenso en los precios desde el año 2014, rozando el 10 %. La zona norte, que recoge municipios como Camas, La Algaba y La Rinconada, es la zona con un precio medio más bajo, unos 414 €.

En cuanto a la disponibilidad de locales comerciales de la zona urbana, durante el año 2015, ésta fue superior en la zona Este, alcanzando el 13,1 % del total de locales, mientras que la zona con menor disponibilidad, como era de esperar, es la zona Centro con una tasa de disponibilidad del 5,3 %.

8.1.3. La ciudad de Sevilla

El municipio de Sevilla se encuentra en el centro de la provincia, concretamente en las coordenadas 37º 23’ norte y 5º 59’ oeste, a una altura de 6 metros sobre el nivel del mar. Es atravesado por el río Guadalquivir. Tiene una superficie de 141,3 kilómetros cuadrados, y representa el mayor núcleo urbano, no sólo de la provincia, también de la comunidad autónoma de Andalucía. (Ayuntamiento de Sevilla, 2016).

Figura 130. Distritos de la ciudad de Sevilla. (Fuente: Ayuntamiento de Sevilla)

Se encuentra dividido en 11 distritos que son Casco Antiguo, Cerro Amate, Norte, Este – Alcosa – Torreblanca, Los Remedios, Macarena, Sur, San Pablo – Santa Justa, Bellavista – La Palmera, Nervión y Triana. Su distribución puede observarse en la Figura 130.

Según la información proporcionada por el INE, a fecha 1 de enero de 2016, tenía una población de 690.566 habitantes, en ligero descenso desde 2010, año en el que la población era de 704.198 habitantes. Esta población representa el 35,6 % de la población total de la provincia.

La densidad de población es, ampliamente superior, a la de la provincia, ya que supera los 4.900 habitantes por kilómetro cuadrado, mientras que la de toda la región es 138,32.

Cuenta con 48.415 empresas en el año 2016, de las que 26.144 corresponden al sector servicios, lo que supone un 54 % del total. Le sigue el comercio, el transporte y la hostelería con 16.733, con un 34,56 % del total. Dentro de las empresas del sector servicios, destacan, con claridad, las destinadas a actividades profesionales y técnicas, que con 11.654 representan un 44,58 % de éste.

Existen 2.754 empresas en la ciudad de Sevilla dedicadas a las actividades inmobiliarias. Éstas, representan un 10,53 % del total de empresas dedicadas al sector servicios, y un 5,69 % del total. Por tanto, en la ciudad, el peso de las empresas con esta

actividad es superior al registrado en la provincia, para la que se obtuvieron porcentajes del 9,52 % y 4,13 % respectivamente.

Analicemos a continuación la información relativa al mercado inmobiliario de la ciudad, con datos de Ministerio de Fomento (2016).

El último dato disponible sobre el número de viviendas de la ciudad de Sevilla, es del censo del año 2011, en el que había un total de 337.225 viviendas, habiendo experimentado un incremento de 40.236 viviendas, respecto al dato anterior correspondiente al año 2001. Del total de viviendas, el 79,6 % se corresponden a viviendas principales. Entre las viviendas no principales, algo más del 70 % se encontraban vacías, un total de 48.178.

El número de edificios destinados principal o exclusivamente a viviendas en este mismo año en la ciudad de Sevilla, ascendía a 56.606 y la distribución de éstos, según estado de conservación y año de construcción, puede consultarse en la Tabla 26.

Tabla 26. Distribución de edificios destinados a vivienda según antigüedad y estado. Sevilla.

Edificios	Ruinoso	Malo	Deficiente	Bueno	Total
Antes de 1900	36	62	159	744	**1.001**
De 1900 a 1920	20	39	148	489	**696**
De 1921 a 1940	40	88	434	2.083	**2.645**
De 1941 a 1950	56	92	427	2.582	**3.157**
De 1951 a 1960	58	270	926	7.652	**8.906**
De 1961 a 1970	57	119	1.078	10.339	**11.593**
De 1971 a 1980	29	92	466	9.022	**9.609**
De 1981 a 1990	6	9	106	6.879	**7.000**
De 1991 a 2001	0	32	117	6.711	**6.860**
De 2002 a 2011	0	11	44	5.084	**5.139**
Total	**302**	**814**	**3.905**	**51.585**	**56.606**

Fuente: INE

Más del 20 % de los edificios de viviendas de la ciudad, en 2011, se construyeron en la década de los 60, y más de la mitad de ellos lo fueron entre 1951 y 1980. Existen un

total de 1.001 edificios de viviendas que son anteriores al siglo XX, y sólo el 3,6 % se encuentran en estado ruinoso.

Tabla 27. Número de viviendas principales en función de la superficie útil y el tamaño del hogar.

	1 pers.	2 pers.	3 pers.	4 pers.	5 pers.	6 o más	Total
Menos de 30m²	435	239					**750**
Entre 30 y 45m²	5.713	3.538	1.593	755	303		**11.939**
Entre 46 y 60m²	16.204	12.042	7.675	4.221	961	661	**41.764**
Entre 61 y 75m²	18.522	20.301	15.980	12.942	2.437	995	**71.177**
Entre 76 y 90m²	13.876	20.222	16.466	16.394	4.091	982	**72.031**
Entre 91 y 105m²	4.714	7.680	5.875	6.428	2.075	529	**27.301**
Entre 106 y 120m²	2.418	4.523	3.893	4.452	1.896	440	**17.622**
Entre 121 y 150m²	2.066	4.032	2.895	3.677	1.654	518	**14.842**
Entre 151 y 180m²	668	1.225	1.136	1.354	724	199	**5.307**
Más de 180m²	700	1.566	1.159	1.282	717	278	**5.702**
Total	**65.317**	**75.368**	**56.717**	**51.535**	**14.859**	**4.640**	**268.435**

Fuente: INE

Por otro lado, la información relativa a la superficie útil en función del número de personas que habitan la vivienda principal, se muestra en la Tabla 27.

Se observa, que más de la mitad de las viviendas principales tiene una superficie útil comprendida entre 61 y 90 m^2, y en estos casos, lo más común es que estén habitadas por 2 personas. Sólo el 4,1 % de las viviendas tienen una superficie mayor de 150 m^2, y un porcentaje similar, el 4,7 % tienen un tamaño por debajo de los 45 m^2.

Por otro lado, según este censo de viviendas, publicado en el año 2011, de las 268.435 viviendas principales de la ciudad, el 45,49 % lo son en propiedad sin hipotecas a cargo, el 31,8 % son en propiedad con cargas y el 11,81 % son alquiladas.

En el tercer trimestre de 2016, se han realizado 1.523 transacciones de viviendas. Al igual que ocurre a nivel provincial, actualmente, este número se encuentra en recuperación tras la disminución experimentada durante los últimos años debido a la crisis inmobiliaria. Los resultados de esta evolución pueden verse en la Figura 131, construida a partir de los datos ofrecidos por el Ministerio de Fomento. El comportamiento de esta

serie a lo largo de los últimos 13 años es similar al observado a nivel provincial, y que se presentó en la Figura 128.

Este volumen de transacciones representa, en los dos últimos años, entre un 45 % y un 48 % del volumen trimestral a nivel provincial, lo que significa que en el municipio de Sevilla se concentra el negocio inmobiliario de la provincia.

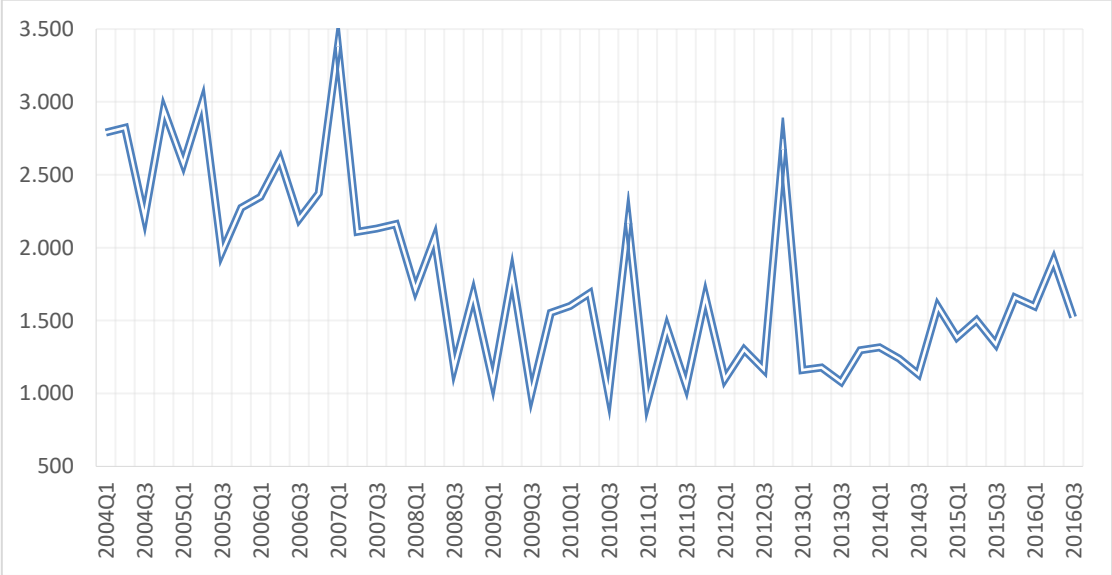


Figura 131. Transacciones de viviendas de la ciudad de Sevilla. (Fuente: Ministerio de Fomento)

De las 1.523 transacciones registradas en este trimestre, 1.289 se corresponden con viviendas libres - un 84,64 % - por lo que sólo un 15,36 % son transacciones correspondientes a viviendas protegidas. Además, 1.421 de las transacciones, lo son de viviendas de segunda mano, por lo que sólo 102 son de nueva construcción.

Una vez analizado el número de transacciones realizadas en la ciudad de Sevilla, veamos el precio en euros por metro cuadrado de éstas. La evolución trimestral de este precio, para el total de viviendas y distinguiendo entre las que tienen más de 5 años de antigüedad o no, y el período que va desde el primer trimestre de 2011 hasta el tercer trimestre de 2016, se muestra en la *Figura 132*.

El precio de las viviendas ha experimentado un descenso paulatino a lo largo de los últimos cinco años. Las viviendas con hasta 5 años de antigüedad, han sufrido una mayor

variabilidad, y la tendencia global es casi coincidente a la de las viviendas con más de 5 años de antigüedad, debido al mayor volumen de transacciones de éstas.

En el tercer trimestre de 2016, el precio medio por metro cuadrado de una vivienda se situó en 1.512,20 €, entre las viviendas de menor antigüedad este precio fue de 1.639,40 € y entre las de mayor, 1.509,40 €. En el primer trimestre de 2013, el precio de la vivienda con hasta cinco años experimentó un descenso superior al 19,5 % respecto al trimestre anterior, acompañado de un descenso del 60 % en el número de transacciones, debido a la medida del gobierno consistente en eliminar la deducción para viviendas compradas a partir de 2013, como se ha mencionado anteriormente.

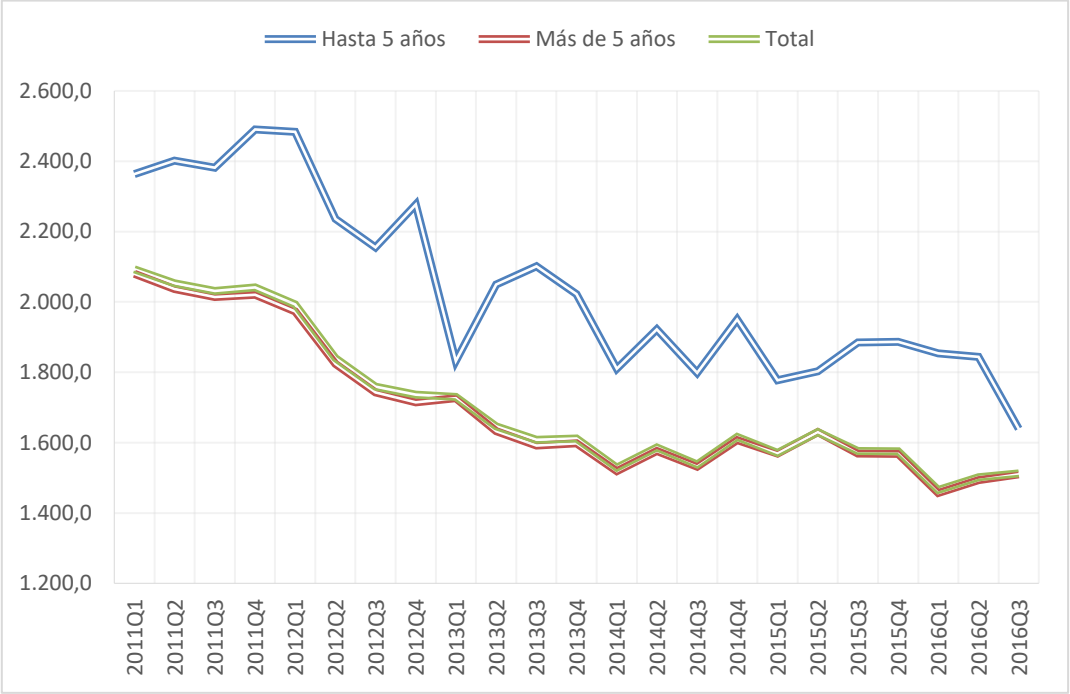


Figura 132. Evolución del precio de la vivienda en la ciudad de Sevilla. (Fuente: Ministerio de Fomento)

Por otro lado, el precio medio por metro cuadrado del suelo urbano del municipio sevillano, se sitúa en el tercer trimestre de 2016, en 219,60 €, experimentando así una reducción del 1,6 % respecto al trimestre anterior, pero un aumento del 71,4 % anual. La evolución de éste entre 2004 y 2016 se muestra en la Figura 133.

El precio del suelo en la ciudad ha sufrido, como en el resto de valores observados, un cambio estructural, creciendo desde 2004 hasta 2008, y decreciendo desde entonces hasta la actualidad.

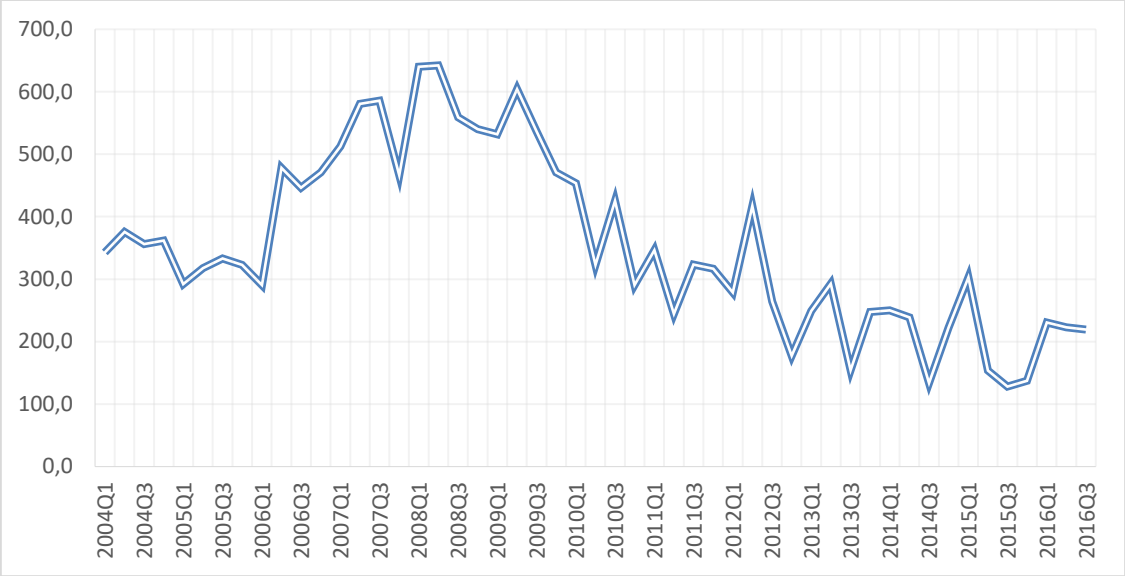


Figura 133. Evolución del precio del suelo en la ciudad de Sevilla. (Fuente: Ministerio de Fomento)

8.2. Análisis del mercado de viviendas a la venta, en la ciudad de Sevilla

Utilizando el programa desarrollado en este trabajo, se ha realizado un análisis estadístico de las viviendas a la venta, en oferta, de la ciudad de Sevilla en la actualidad, ofertadas en el portal estadístico www.idealista.com Este análisis tiene como objetivo obtener una visión de la oferta de viviendas a la venta de la ciudad junto con las características que las definen, identificando viviendas que puedan tener características diferenciadoras del resto y obtener un modelo que nos permita estimar el precio de oferta de una vivienda, a partir de las características que la definen.

Para cada vivienda, se han recogido las siguientes variables:

- Longitud del inmueble
- Latitud del inmueble
- Distancia al centro de la búsqueda.
- Barrio al que pertenece
- Distrito al que pertenece
- Municipio al que pertenece. Restringiremos el estudio a la ciudad de Sevilla, por lo que esta variable será usada únicamente como filtro.
- Código del inmueble

- Tipo de vivienda
- Número de habitaciones
- Número de baños
- Tamaño (en metros cuadrados)
- Precio (en euros)
- Planta en la que se encuentra situado el inmueble
- Dispone / no dispone de cochera
- Dispone / no dispone de ascensor
- Dispone / no dispone de piscina
- Dispone / no dispone de aire acondicionado
- Dispone / no dispone de terraza
- Dispone / no dispone de trastero
- Dispone / no dispone de armarios empotrados
- Estado de conservación de la vivienda
- Número de fotos del anuncio

La búsqueda de viviendas a la venta se realizó el día 1 de enero de 2017, tomando, como centro de la búsqueda, el punto de coordenadas de latitud 37˚ 23’ norte y longitud 5˚ 59’ oeste, correspondiente a la intersección entre la calle San Bernardo y la calle Cofia del barrio de San Bernardo, a escasos metros del edificio de la Diputación de Sevilla. Desde este punto, se ha considerado un radio de 8.000 metros, que abarca todo el municipio, así como algunos municipios del extrarradio, como son Tomares, San Juan de Aznalfarache, Dos Hermanas o Montequinto.

8.2.1. Análisis descriptivo

Se ha obtenido información de un total de 12.951 viviendas a la venta, de las que 10.083 se encuentran situadas en el término municipal de Sevilla. Estas serán las viviendas que serán analizadas. Cabe destacar que los resultados mostrados han sido procesados por el programa propuesto en este trabajo.

Las viviendas a la venta por un menor precio están situadas en el distrito Cerro Amate, en una zona deprimida de la ciudad. Son dos pisos de 70 y 80 metros cuadrados por un precio de venta de tan sólo 10.000 €.

La vivienda de mayor precio está situada en la zona centro de la ciudad, cerca de Plaza Nueva y la calle Tetuán. Es un chalé pareado de casi 2000 metros cuadrados, distribuidos en tres plantas, y que tiene un precio de venta de 7.100.000 €.

La vivienda a la venta de menor tamaño, es un estudio con una superficie de 18 metros cuadrados, situado en la entreplanta de un bloque de viviendas del barrio de Alameda.

Por el contrario, la vivienda de mayor tamaño ofertada es una casa de campo, tipo cortijo, situada en Avenida de las Ciencias del distrito Sevilla Este. Tiene una superficie construida de 3.200 metros cuadrados, en una finca de casi 8.000 metros cuadrados.

Desde el punto de vista del anuncio, el que más fotografías del inmueble ofrece, es una oferta en el que vende un piso situado en el barrio de la Encarnación del distrito Centro, del que se muestran un total de 58 fotografías.

Veamos a continuación, un resumen descriptivo de las principales variables analizadas. Comenzaremos por analizar el precio de la vivienda, cuyo resumen se muestra en la Tabla 28.

Tabla 28. Estudio descriptivo. Precio. Sevilla. (Fuente: Elaboración propia)

Valores	Precio
Media	274363,5829
Mediana	193000,0000
Cuasidesviación típica	284936,1201
Cuasivarianza	81188592534,1691
Rango	7090000,0000
Coeficiente de variación	1,0385
Coeficiente de asimetría	4,7376
Coeficiente de apuntamiento	51,7941
Ext. inf. IC media 95 %	268800,7509
Ext. sup. IC media 95 %	279926,4148
Percentil 5	53000,0000
Percentil 25	112000,0000
Percentil 75	340000,0000
Percentil 95	750000,0000

Fuente: INE

Como puede observarse, el precio medio de las viviendas a la venta en la ciudad de Sevilla es de 274.363, 58 €, con una alta variabilidad y con numerosas observaciones atípicas. Con un 95 % de confianza, podemos afirmar, además, que el precio está comprendido entre 268.800,75 € y 279.926,41 €. El 50 % de las viviendas tiene un precio

de venta inferior a 193.000 € y también la mitad de ellas, tiene un precio comprendido entre 112.000 € y 340.000 €.

El diagrama de caja de la Figura 134, nos confirma la existencia de los numerosos inmuebles con precios que pueden considerarse atípicos superiores. Destacar que se ha eliminado el inmueble de 7.100.000 € de precio, para mejorar la visibilidad del diagrama.

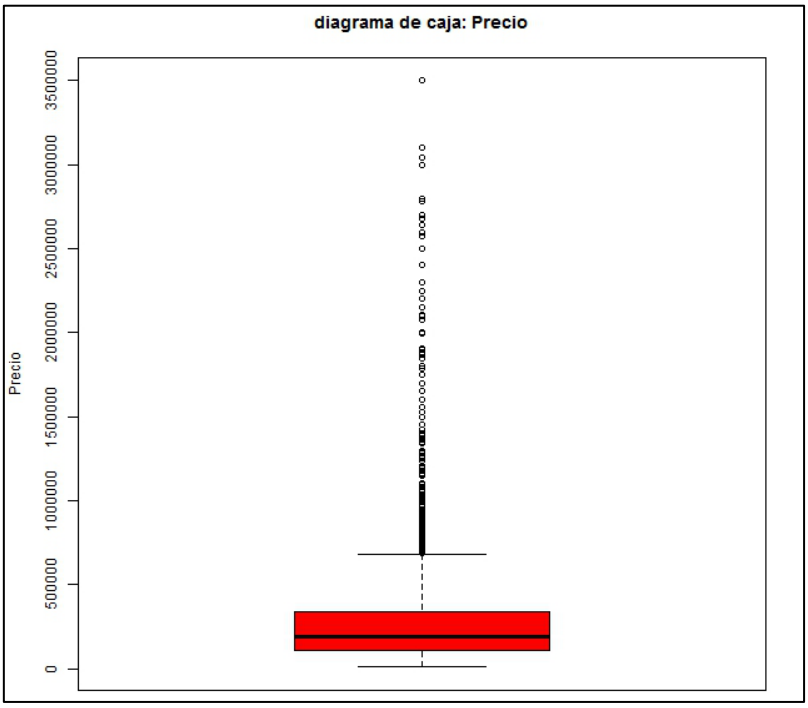


Figura 134. Diagrama de caja. Precio. Sevilla. (Fuente: Elaboración propia)

La distribución geográfica de las viviendas, en oferta, en el municipio de Sevilla, se muestra en la Figura 135. En ésta, se muestra, además, con un degradado de color que va desde el amarillo, para los inmuebles de menor precio, hasta el rojo, para los de mayor precio. Se han eliminado las viviendas con precio superior al millón de euros para una mejor visualización de la escala de color.

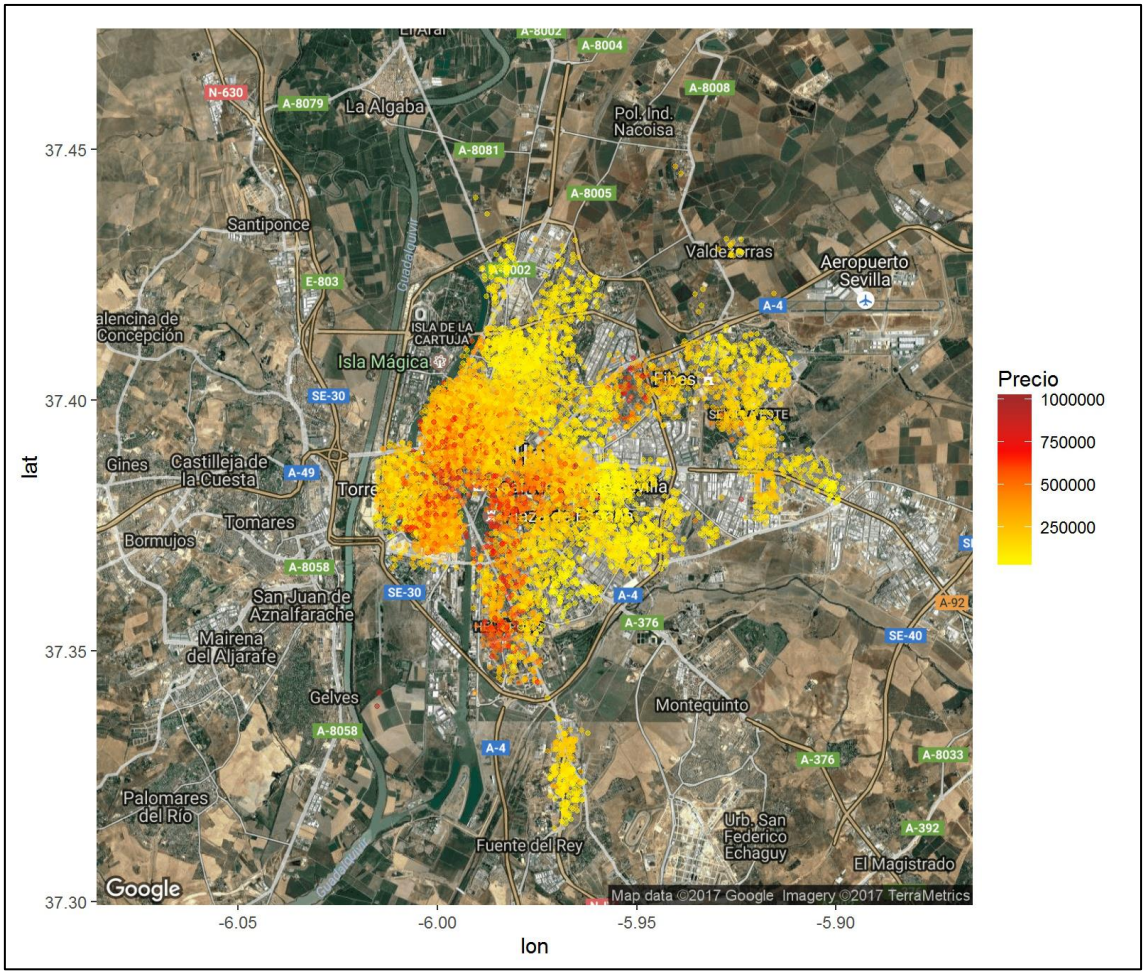


Figura 135. Distribución geográfica. Precio. Sevilla. (Fuente: Elaboración propia)

Como podemos observar, los inmuebles de mayor precio se concentran en la zona centro, en el barrio de los Remedios, La Palmera, Heliópolis, Santa Clara y Nervión.

Analicemos a continuación el tamaño de los inmuebles estudiados. La Tabla 29 muestra el resultado del estudio descriptivo. El tamaño medio de las viviendas del municipio es de 131,96 metros cuadrados, y su valor mediano es 100 metros cuadrados.

Existe una menor dispersión en el tamaño de las viviendas, pero siguen existiendo numerosos valores atípicos. Como se puede ver, el 90 % de los inmuebles tienen una superficie comprendida entre 54 y 309 metros cuadrados, mientras que el rango de la variable es 3182 metros cuadrados. La variable es claramente leptocúrtica y asimétrica a la derecha.

Tabla 29. Estudio descriptivo. Tamaño. Sevilla.

Valores	Tamaño
Media	131,9647
Mediana	100,0000
Cuasidesviación típica	112,4030
Cuasivarianza	12634,4297
Rango	3182,0000
Coeficiente de variación	0,8518
Coeficiente de asimetría	9,0561
Coeficiente de apuntamiento	180,1065
Ext. inf. IC media 95 %	129,7702
Ext. sup. IC media 95 %	134,1591
Percentil 5	54,0000
Percentil 10	61,0000
Percentil 25	76,0000
Percentil 75	150,0000
Percentil 90	240,0000
Percentil 95	309,0000

Fuente: Elaboración propia

La Figura 136, muestra el mapa geográfico con la distribución de los inmuebles según su tamaño. La escala se ha cambiado para mostrar la posición más de cerca.

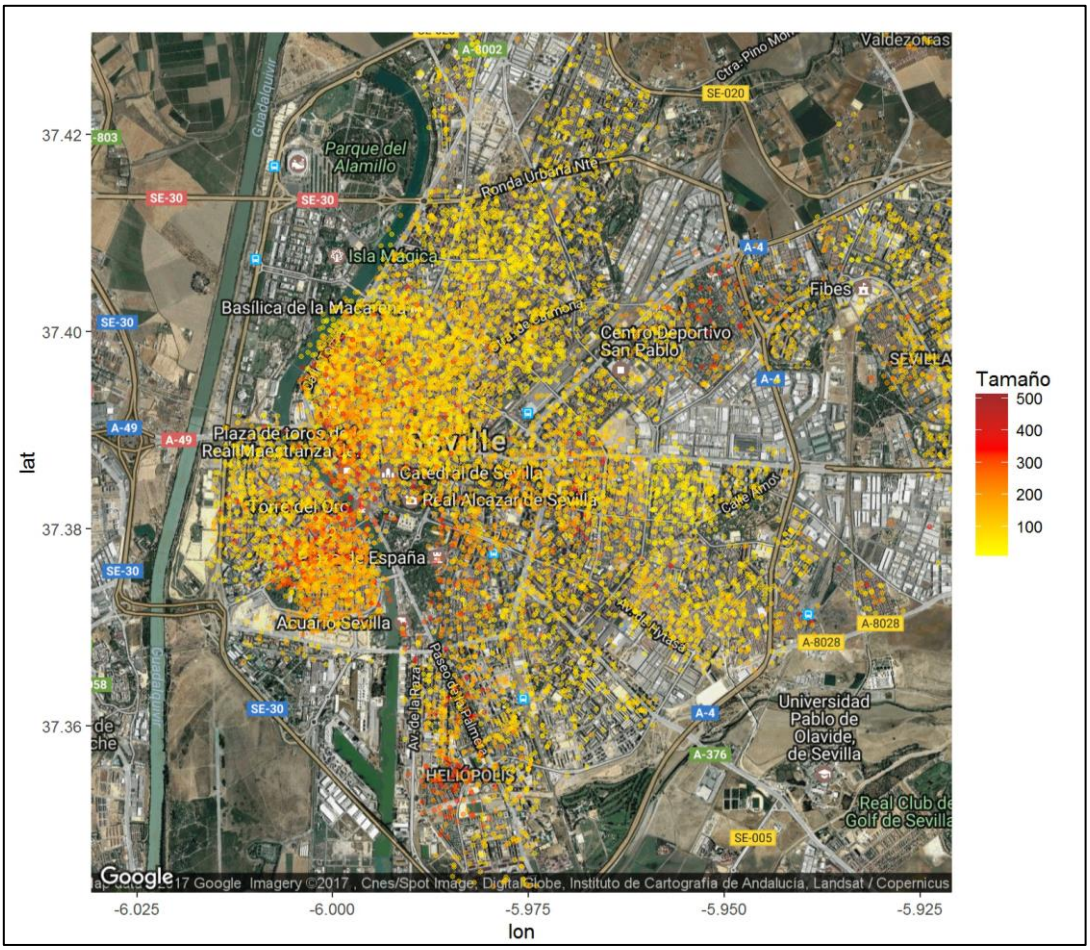


Figura 136. Distribución geográfica. Tamaño. Sevilla. (Fuente: Elaboración propia)

La situación de las viviendas de mayor tamaño coincide, en buena parte, a la de las de mayor precio. Cabe destacar la concentración de viviendas de gran tamaño en Heliópolis y Santa Clara, zonas en las que los predominan las construcciones de vivienda unifamiliar como dúplex, chalés y casas rurales.

Para desligar el precio de la vivienda con su tamaño, analizaremos, a continuación, el precio por metro cuadrado de éstas. La Tabla 30, muestra los resultados del estudio descriptivo realizado para esta variable.

Tabla 30. Estudio descriptivo. Precio por metro cuadrado. Sevilla.

Valores	Precio / m^2
Media	1971,4949
Mediana	1900,0000
Cuasidesviación típica	920,4250
Cuasivarianza	847182,0929
Rango	8698,5294
Coeficiente de variación	0,4669
Coeficiente de asimetría	0,9821
Coeficiente de apuntamiento	2,6143
Ext. inf. IC media 95 %	1953,5271
Ext. sup. IC media 95 %	1989,4626
Percentil 5	716,3333
Percentil 10	885,9001
Percentil 25	1271,5789
Percentil 75	2521,3675
Percentil 90	3124,6667
Percentil 95	3545,4545

Fuente: Elaboración propia

El precio medio de oferta por metro cuadrado, de las viviendas analizadas, es de 1971,49 €, por encima del valor registrado por el INE, indicado en la Figura 132, que ascendía en el segundo trimestre de 2016 a 1843 € y a 1639 € en el tercero. El valor obtenido es compatible con los dados por el INE si tenemos en cuenta que los precios aquí analizados son precios de oferta y no de venta, por lo que éste puede sufrir variación en la negociación de la compra – venta del inmueble.

Los inmuebles de menor precio por metro cuadrado coinciden con los de menor precio total, antes comentados. El de mayor precio es un ático de 170 metros cuadrados, situado en Prado de San Sebastián que tiene un precio de venta de 1.500.000 €. No obstante, el 95 % de las viviendas a la venta tienen un precio por metro cuadrado inferior a los 3.545 € y sólo el 5 % de éstas tiene un precio por debajo de los 716 €.

Para analizar la distribución sobre la ciudad del precio por metro cuadrado, se ha eliminado el 1 % de las viviendas en oferta con menor y mayor precio por metro cuadrado, por lo que la Figura 137, muestra la posición de las viviendas con precios por metro cuadrado comprendidos entre 450 € y 4.600 €.

Como cabía esperar, las viviendas con mayor precio por metro cuadrado se encuentran situadas en el casco histórico, Prado de San Sebastián, Nervión, Los Remedios y Triana.

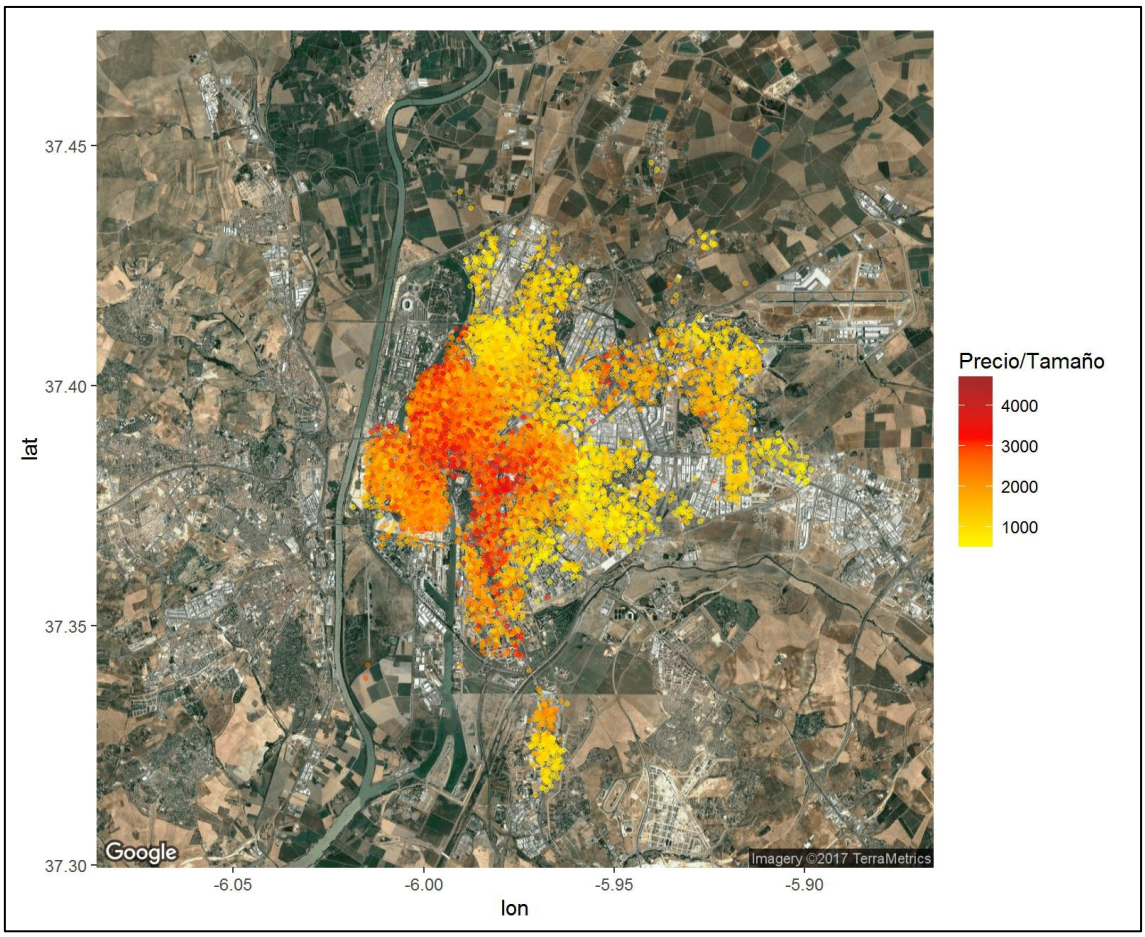


Figura 137. Distribución geográfica. Precio por metro cuadrado. Sevilla. (Fuente: Elaboración propia)

8.2.2. Análisis por distritos

La distribución del precio de la vivienda por metro cuadrado difiere significativamente en los distintos distritos de la ciudad de Sevilla. Los resultados obtenidos al profundizar en el análisis de esta característica, en cada zona de la ciudad se muestran a continuación. Para ello se ha dividido la ciudad en 17 zonas.

8.2.2.1. Distrito de Bellavista – Jardines de Hércules

Es la zona de la ciudad más al sur, y se encuentra en el extrarradio de ésta. El precio medio del metro cuadrado es muy inferior al de la ciudad, con 1306,39 €, sobre un total de 256 inmuebles analizados. Se distinguen dos zonas, como puede observarse en la Figura 138. En la zona norte se encuentra el barrio de los Jardines de Hércules, donde el precio por metro cuadrado es de 1976, 42 €, mientras, que, en la zona sur, se encuentra el barrio de Bellavista, con un precio de 1068,87 €/m^2.

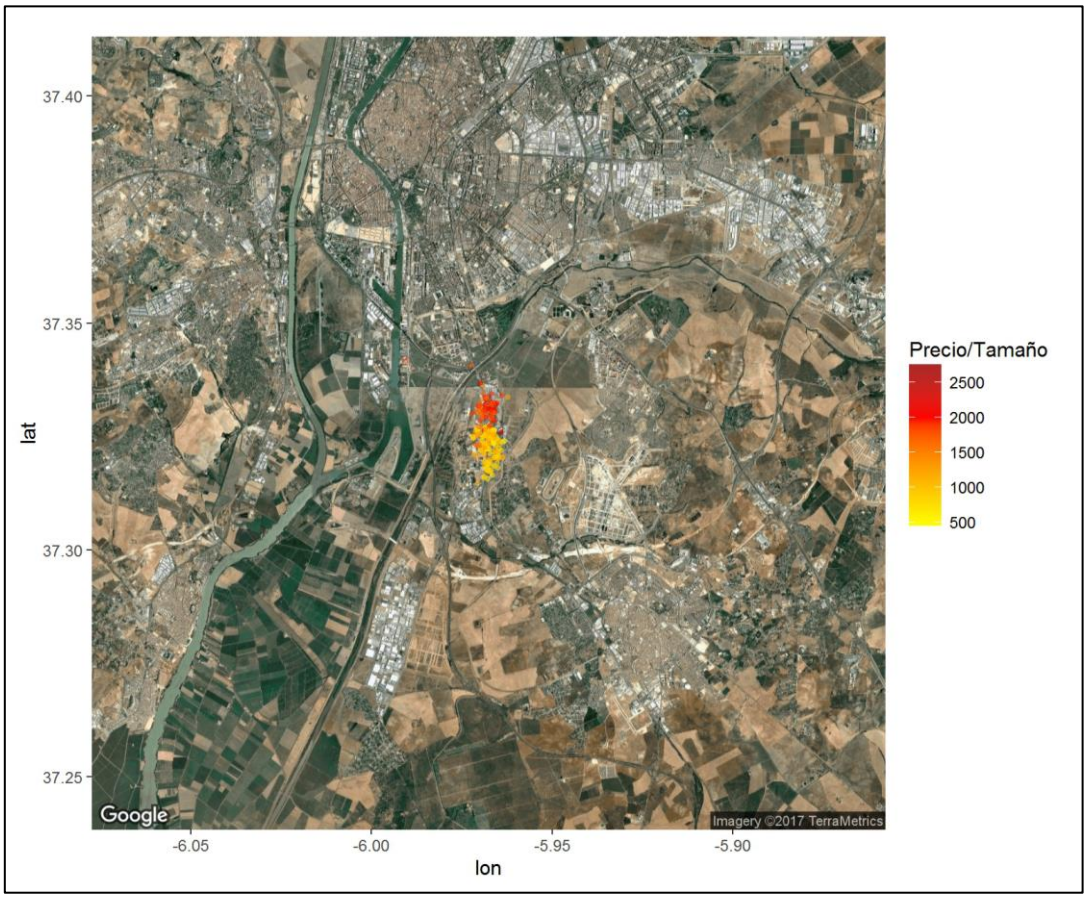


Figura 138. Distribución geográfica. Precio por metro cuadrado. Bellavista – Jardines de Hércules. (Fuente: Elaboración propia)

8.2.2.2. Distrito Centro

Se encuentra en la orilla este del río Guadalquivir. El precio por metro cuadrado es el más alto de la ciudad, con 2.684 €. Este valor, se ha calculado en base a una muestra de 2.467 viviendas. Se distinguen, a su vez, 10 zonas: Alameda, Arenal – Museo,

Encarnación – Las Setas, Feria, Gavidia – San Lorenzo, Puerta Carmona - Puerta Osario – Amador de los Ríos, Puerta de la Carne – Judería, San Julián, San Vicente y Santa Cruz – Alfalfa.

Las viviendas situadas en el barrio de Santa Cruz – Alfalfa son las de mayor precio medio, con 3.181 €, seguido de Arenal – Museo, con casi 2.900 € / m^2. Por el contrario, el barrio de San Julián, al nordeste del distrito, es el que presenta e precio más bajo, con 2.173 € / m^2.

La distribución de viviendas y la situación del barrio en el municipio, se muestran en la Figura 139.

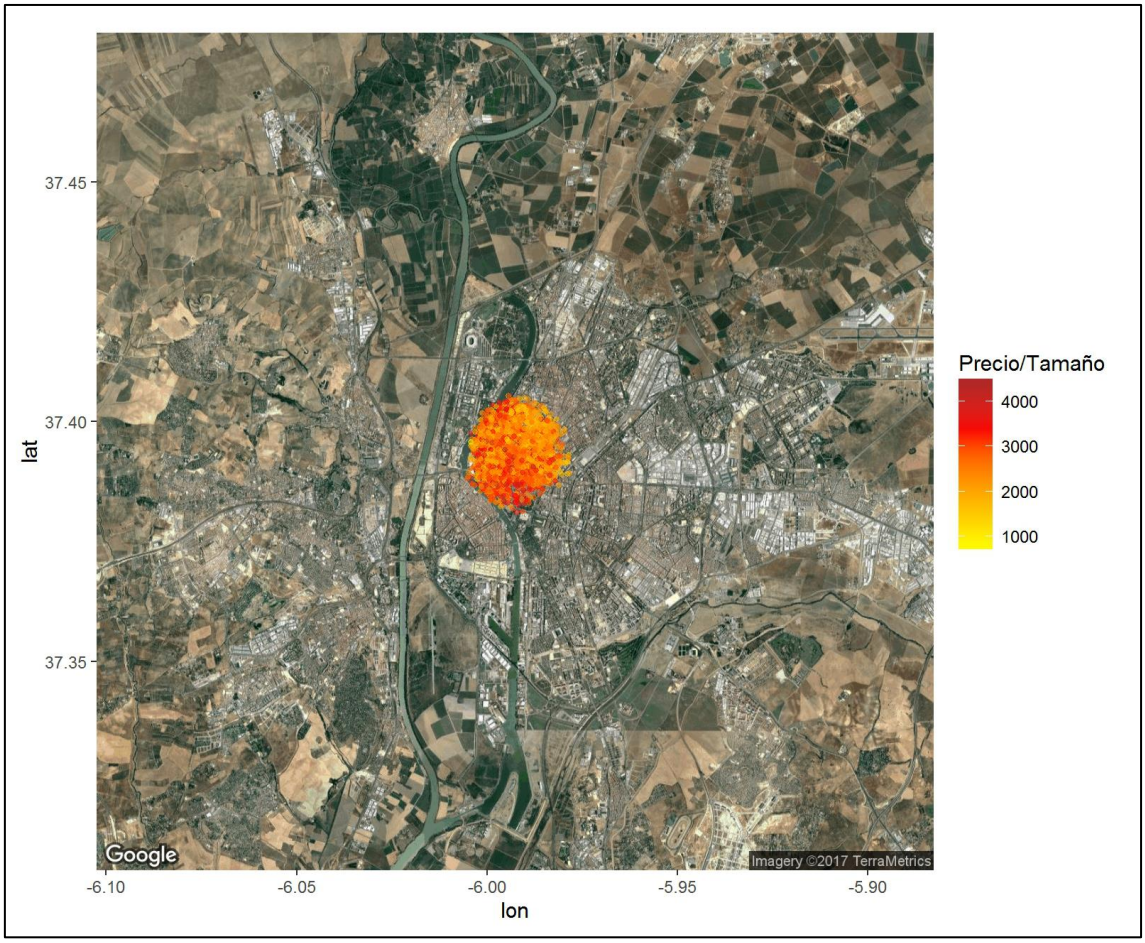


Figura 139. Distribución geográfica. Precio por metro cuadrado. Centro. (Fuente: Elaboración propia)

8.2.2.3. Distrito Cerro - Amate

Esta zona está enclavada en el sureste de la ciudad de Sevilla. En el momento de la búsqueda se hallaron un total de 1.110 viviendas a la venta, cuyo precio era de 960,45 euros por metros cuadrados. Pueden distinguirse, también, 10 barrios en su interior: Amate, Cerro del Águila, Juan XXIII, La Plata, Los Pajaritos, Palmete - Padre Pío - Hacienda San Antonio, Rochelambert, Santa Aurelia y Su Eminencia – La Oliva.

El barrio con el precio por metro cuadrado más elevado es el de Santa Aurelia, en la zona nordeste del distrito. Su precio: 1.207,43 € / m^2. Sólo dos barrios más superan los 1.000 € por metro cuadrado: Palmete – Padre Pío – Hacienda San Antonio y Cerro del Águila. En el lado opuesto se encuentra el barrio de Los Pajaritos, con 611,90 € / m^2. No en vano, es uno de los barrios con menor renta per cápita de España, y una de las zonas, junto con Cerro del Águila y Su Eminencia – Oliva, más deprimidas de la ciudad. El barrio de La Oliva, junto al polígono Sur, zona conocida como las Tres Mil Viviendas, pertenecen administrativamente al distrito Sur, aunque han sido consideradas en este estudio en este distrito, por cercanía geográfica y por sus similitudes en cuanto a condiciones socioculturales y económicas.

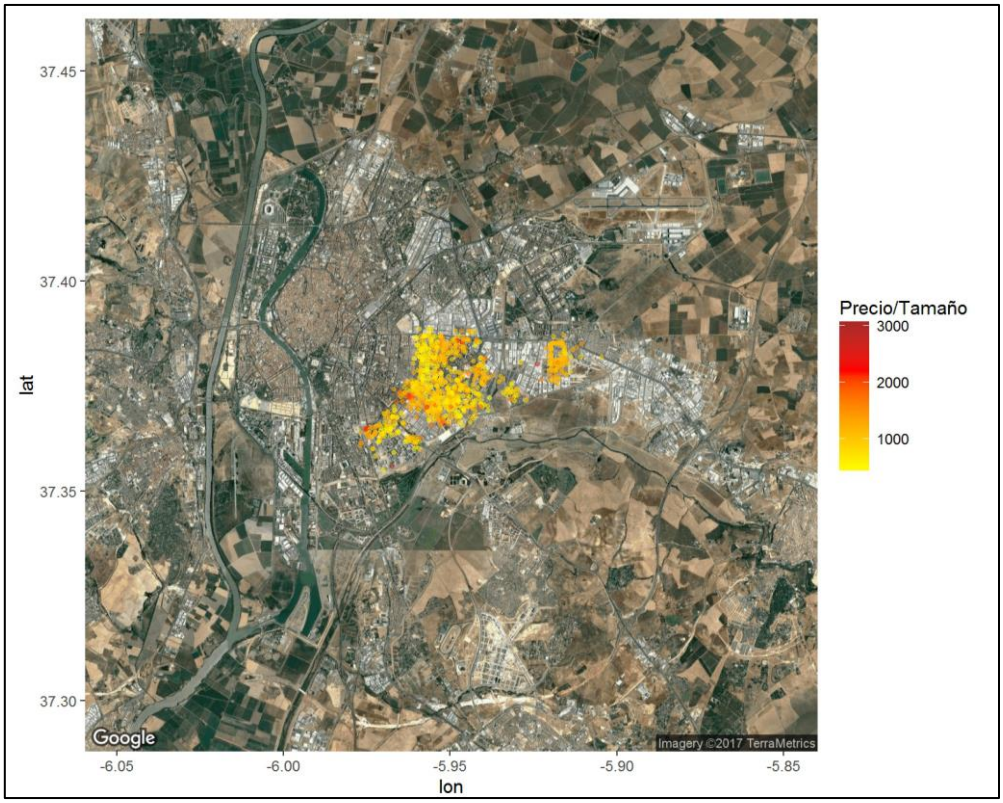


Figura 140. Distribución geográfica. Precio por metro cuadrado. Cerro Amate

La situación de los inmuebles en el distrito y su precio por metro cuadrado puede observarse en la Figura 140. Cabe destacar que no han sido mostradas viviendas en el mapa en la zona del polígono Sur - extremo suroeste - debido a que se ha situado el límite inferior del precio en 450 € / m^2, como se ha mencionado anteriormente.

8.2.2.4. Distrito La Palmera – Los Bermejales

Esta zona se encuentra situada al sur de la ciudad, pero en el interior del núcleo urbano, como puede apreciarse en la Figura 141. En ella, se ha recogido información de un total de 513 viviendas a la venta. Pueden distinguirse 4 barrios en esta zona: Bami – Pineda, La Palmera – Manuel Siurot, Los Bermejales y Reina Mercedes – Heliópolis. En Este último se encuentra uno de los campus de la Universidad de Sevilla.

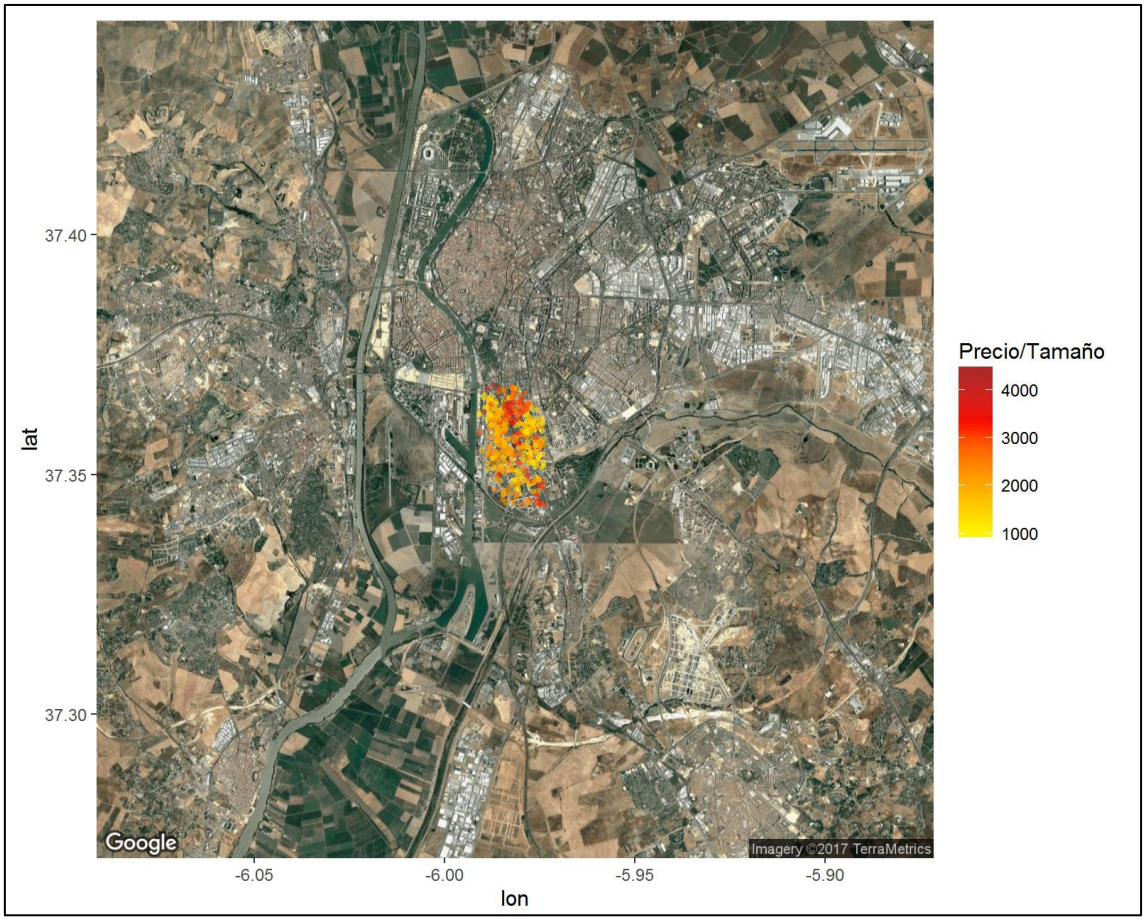


Figura 141. Distribución geográfica. Precio por metro cuadrado. La Palmera – Los Bermejales. (Fuente: Elaboración propia)

El barrio con las viviendas más caras, con una diferencia significativa con respecto a los barrios de la misma zona, se encuentra en La Palmera, con un precio de 2.912,24 € / m^2. Puede distinguirse en la *Figura 141*, como los puntos de esa zona, la parte nordeste, tienen un color rojo más intenso.

Los Bermejales y Reina Mercedes – Heliópolis tienen precios medios similares, alrededor de los 2.200 € / m^2. Bami – Pineda dispone del precio medio más bajo: 1.878,32 € / m^2.

8.2.2.5. Distrito Los Remedios

Este es uno de los distritos más emblemáticos de la ciudad. Se encuentra situado en la orilla occidental del río Guadalquivir, al sur del barrio de Triana. Se han recogido datos relativos a 890 viviendas a la venta, con un precio medio de 2.292,41 € / m^2. Es, además,

el distrito de la ciudad, seguido de Santa Clara, en el que el precio medio total de la vivienda es mayor, con 426.657 €. (Véase *Figura 142*)

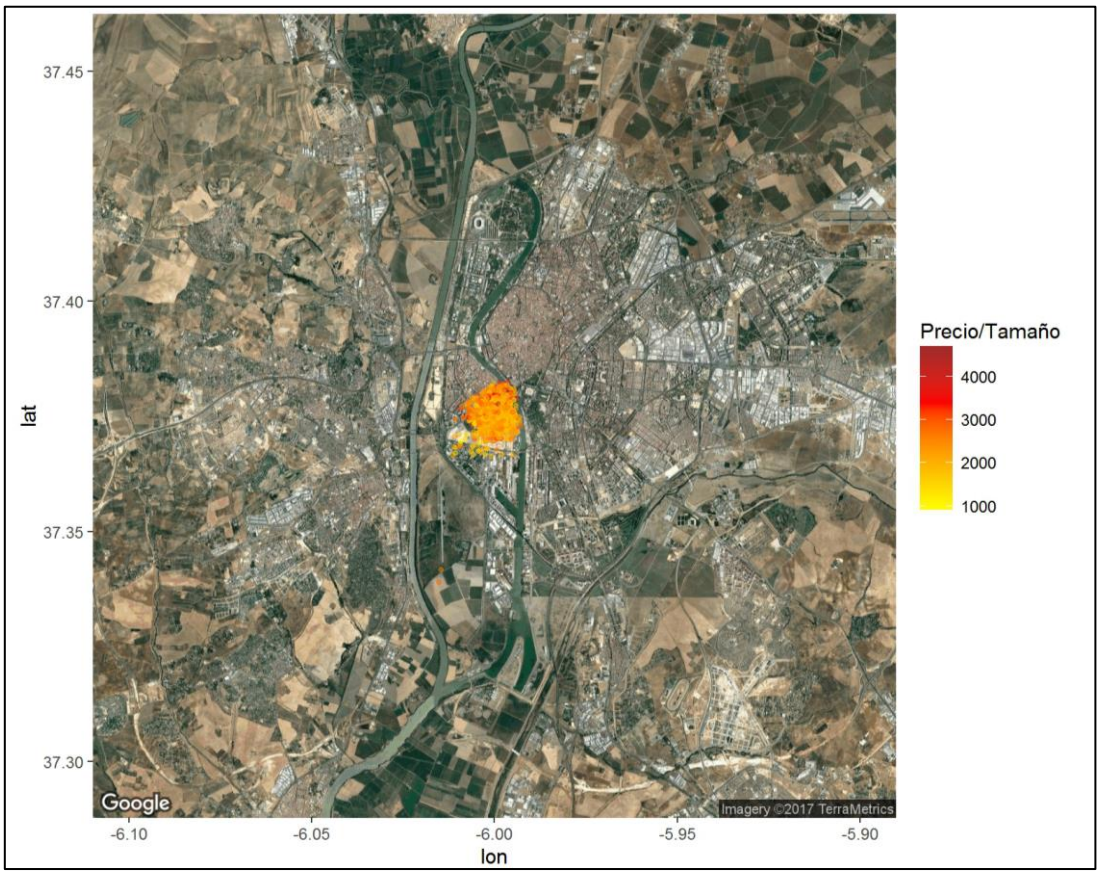


Figura 142. Distribución geográfica. Precio por metro cuadrado. Los Remedios

Podemos distinguir cinco zonas: Asunción – Carrero Blanco, Blas Infante, Parque de los Príncipes – Calle Niebla, Plaza de Cuba – República Argentina y Tablada – Ramón de Carranza – Madre Rafols.

Las dos zonas colindantes al norte con el barrio de Triana, Blas Infante y Plaza de Cuba – República Argentina, son las que presentan un precio medio por metro cuadrado más elevado, en torno a los 2.500 €. El precio medo del resto de zonas baja hasta los 2.200 € / m^2, aún muy por encima del precio medio del metro cuadrado de la ciudad.

8.2.2.6. Distrito Macarena

Este distrito se encuentra, junto a Pino Montano, en la zona norte de la ciudad. Limita por el sur con el distrito Centro, lo que provoca que el precio medio de la zona Sur del distrito sea significativamente mayor que en el resto de zonas del amplio distrito.

El precio medio, para 711 viviendas analizadas, es de 1.290,48 €, muy por debajo de los más de 1.900 € de la ciudad. Pueden distinguirse cinco barrios o zonas, que en orden creciente a la distancia al centro de la ciudad son: Parlamento – Torneo, Doctor Fedriani, Pío XII, Villegas – Los Príncipes y Los Carteros – San Diego.

El barrio de Parlamento – Torneo, colindante con el distrito Centro y donde se encuentra el edificio del Parlamento de la Comunidad, es el barrio con el precio medio por metro cuadrado más caro, con una gran diferencia con el resto, su precio medio: 2.014,60 € / m^2.

Las cuatro zonas restantes tienen precios medios de entre 1.000 y 1.200 euros por metro cuadrado, lo que las sitúa muy por debajo de la media de la ciudad.

En la *Figura 143* puede distinguirse con claridad, por el distinto color de los puntos que representan las viviendas, las pertenecientes al barrio del Parlamento y Torneo.

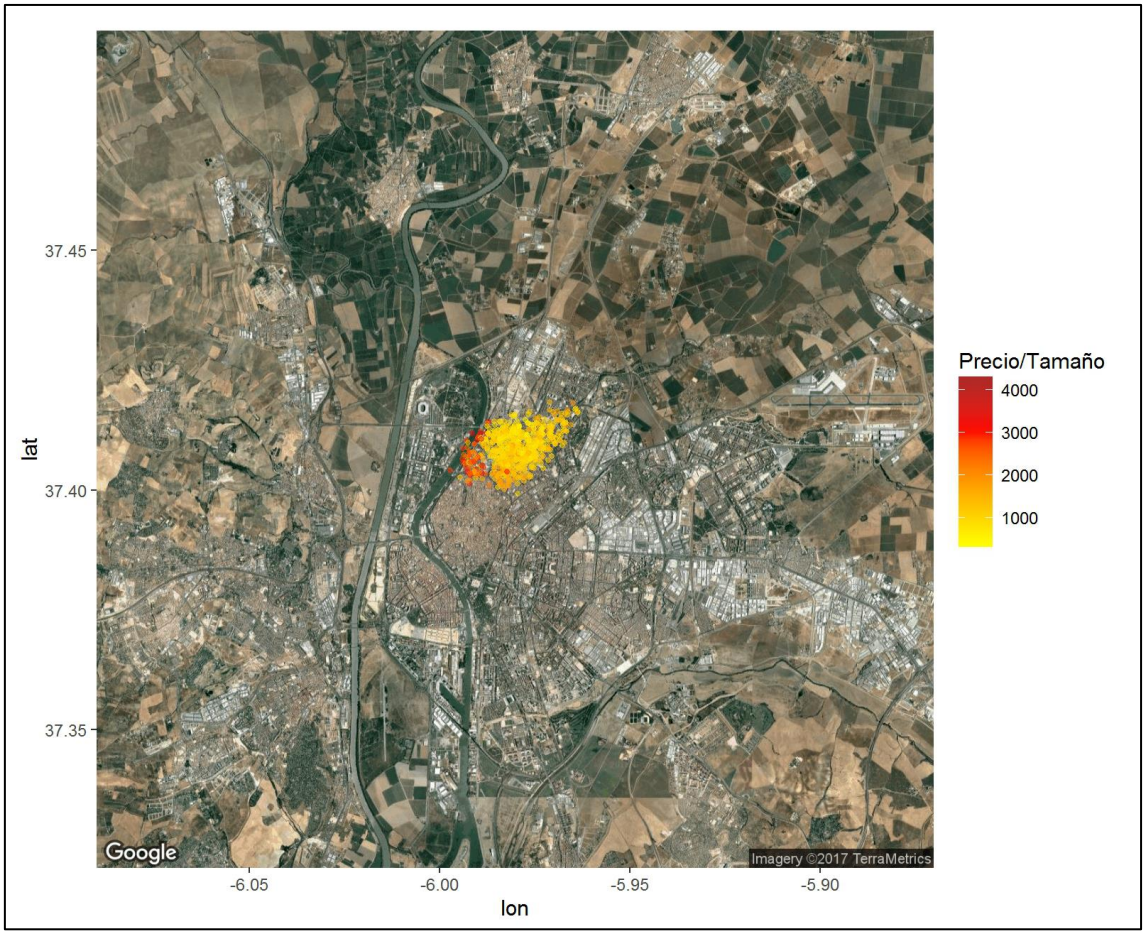


Figura 143. Distribución geográfica. Precio por metro cuadrado. Macarena. (Fuente: Elaboración propia)

8.2.2.7. Distrito Nervión

El distrito de Nervión está situado en el lado oriental de la ciudad. Al este del distrito Centro. La muestra analizada, se compone de 1.108 viviendas, que tienen un precio medio de venta de 2.336,03 €, - por encima del valor medio de la ciudad. - Es, de hecho, el tercer distrito en el que el precio medio por metro cuadrado es mayor, tras Centro y Prado de San Sebastián – Felipe II.

Al igual que en el distrito Macarena, la zona colindante con el distrito Centro tiene un precio medio superior al resto. Podemos distinguir 4 zonas en el interior del distrito: Gran Plaza – Marqués de Pickman – Ramón y Cajal, Luis Montoto – Santa Justa, Nervión y San Bernardo – Buhaira – Huerta del Rey. Este último linda con el distrito Centro y con el parque María Luisa, y es en el que se encontró un mayor precio medio: 2.917,25 € /

m^2. De hecho, es, también, la zona con mayor precio medio tras el barrio Santa Cruz – Alfalfa, del distrito Centro, y los barrios de El Porvenir y el Prado de San Sebastián del distrito Sur.

En la Figura 144 se muestra la distribución de las viviendas en este distrito:

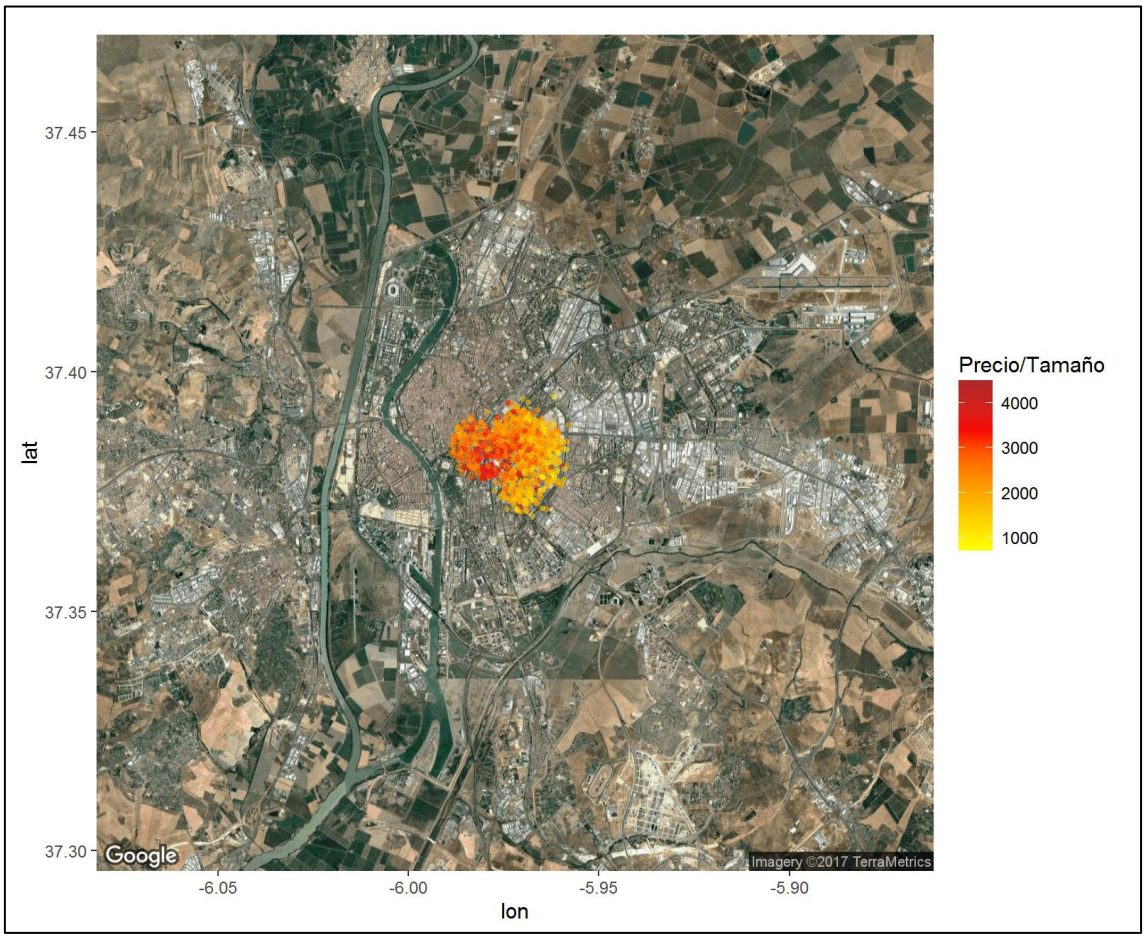


Figura 144. Distribución geográfica. Precio por metro cuadrado. Nervión. (Fuente: Elaboración propia)

8.2.2.8. Distrito Parque Alcosa

Es una zona situada al nordeste de la ciudad de Sevilla, situada a las afueras de ésta y cercana al aeropuerto. Delimita al oeste con San Pablo, y al sur con Sevilla Este. Administrativamente, pertenece al mismo distrito que Sevilla Este y Torreblanca.

En el momento de realizar la búsqueda, había un total de 104 viviendas a la venta, con un precio medio de venta muy inferior al registrado a nivel municipal: 1.066,52 euros

por metro cuadrado, por lo que se sitúa como la cuarta zona de la ciudad con el precio medio por metro cuadrado más bajo tras Torreblanca, Cerro Amate y San Jerónimo.

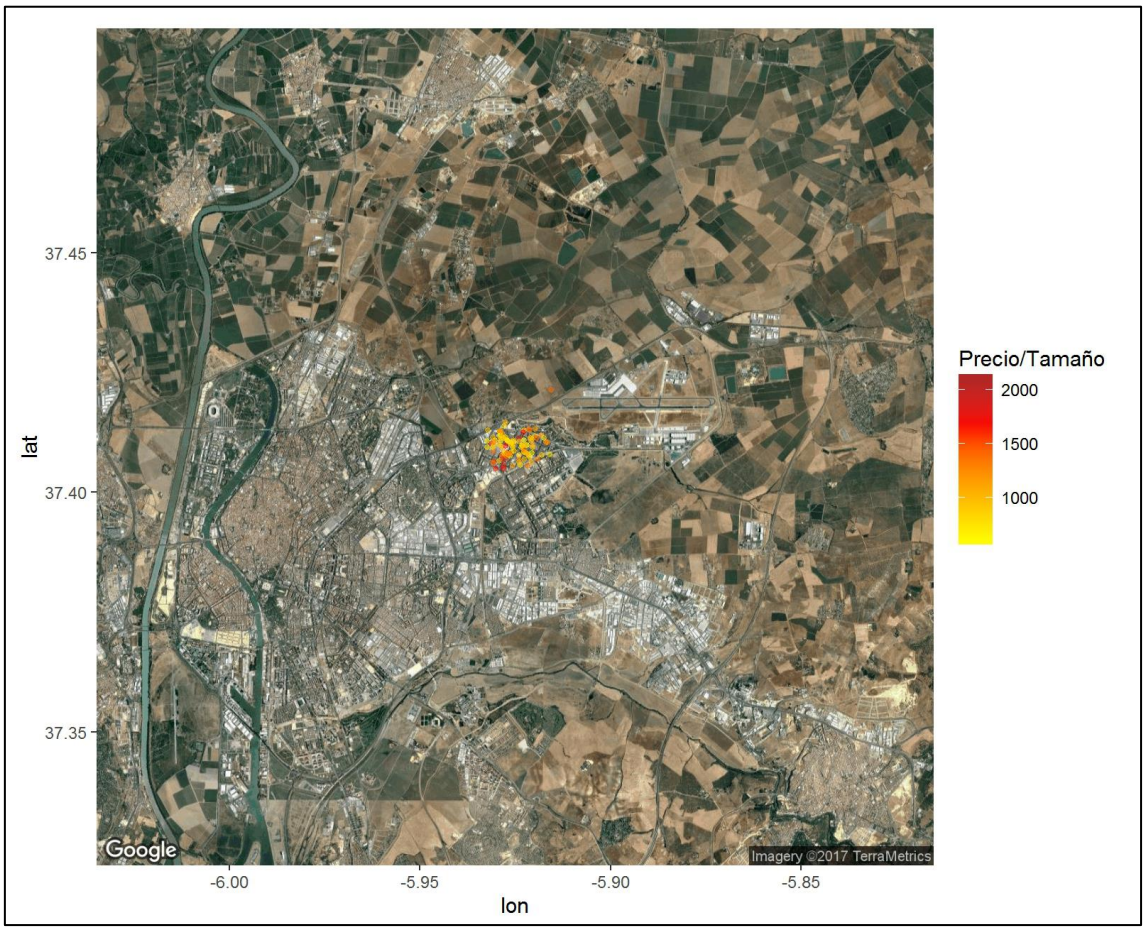


Figura 145. Distribución geográfica. Precio por metro cuadrado. Parque Alcosa. (Fuente: Elaboración propia)

8.2.2.9. Distrito Pino Montano

Esta zona está situada en el extremo norte de la ciudad. Pertenece, administrativamente, al distrito Norte. La muestra analizada contenía 202 viviendas a la venta. Su precio medio se sitúa algo por encima que Parque Alcosa, pero al ser un barrio periférico de la ciudad, éste, 1.148,08 € sigue siendo muy inferior al de la ciudad, que se sitúa alrededor de los 2.000 € / m^2.

Como puede observarse en la Figura 146, a esta zona aparecen asociados algunas viviendas situadas en el extrarradio de la ciudad, aunque dentro del municipio sevillano.

Estas viviendas forman el núcleo urbano de Valdezorras, en el que predominan las casas de campo y los chalés.

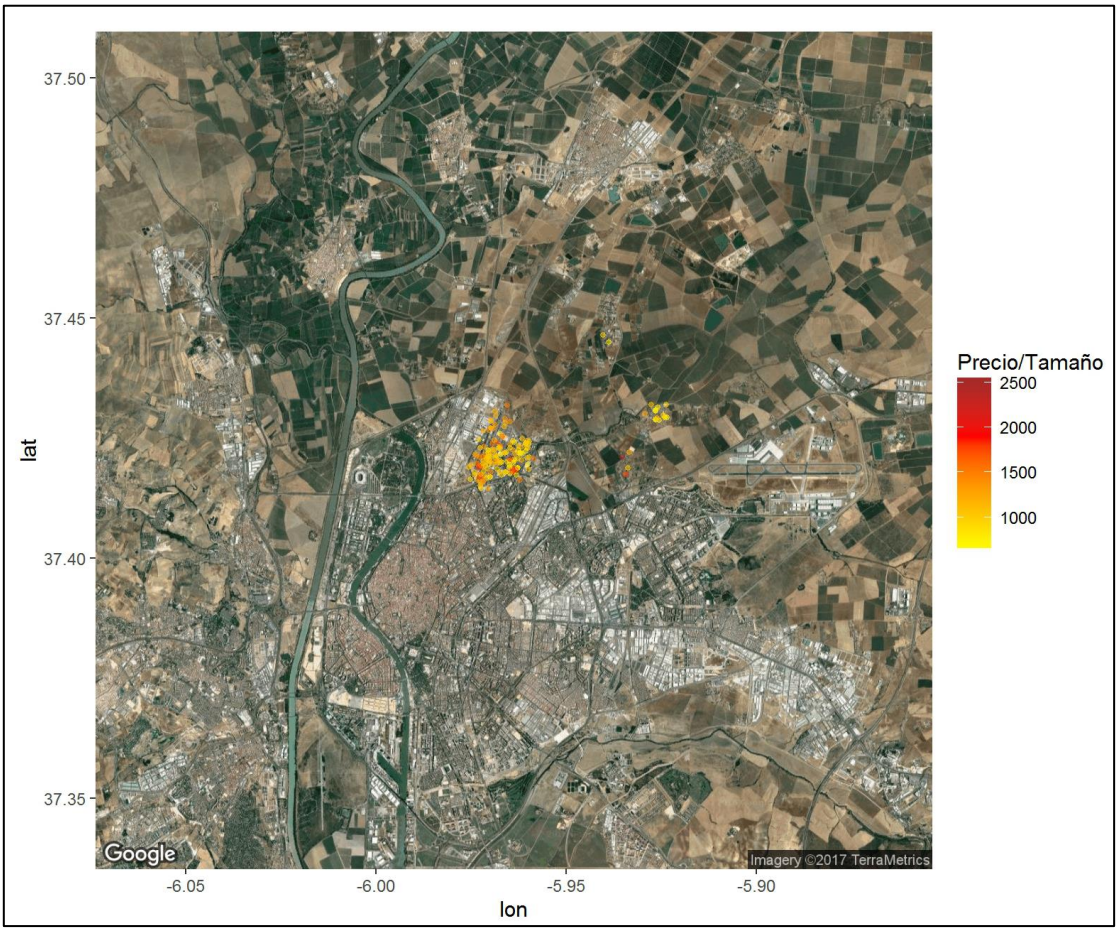


Figura 146. Distribución geográfica. Precio por metro cuadrado. Pino Montano. (Fuente: Elaboración propia)

8.2.2.10. Distrito Prado de San Sebastián – Felipe II

Esta zona, perteneciente administrativamente al distrito Sur, se encuentra situada al sur del barrio de Nervión y al oeste de Cerro Amate. Pueden distinguirse cuatro barrios que son El Porvenir, Felipe II – Bueno Monreal, Prado de San Sebastián y Tiro de Línea.

Se han analizado 353 viviendas, que han devuelto un precio medio de 2.480,36 € / m^2, la segunda zona en la que hemos dividido el municipio, con el mayor valor, tras el distrito Centro.

De los cuatro barrios en los que se divide, los dos primeros tienen precios medios cercanos a los 3.000 €, de hecho, son dos de las zonas con precios más caros de la ciudad. En el barrio de Felipe II – Bueno Monreal, también se han encontrado un precio medio por encima de los 2.500 € / m^2. Destaca en el otro extremo el barrio de Tiro de Línea, situado en el extremo sureste, ya que, en éste, el precio medio es de tan sólo 1.475,20 €. Esto lo justifica el hecho de que linda sur con la barriada conocida como Las Tres Mil Viviendas, zona deprimida de la ciudad.

En la Figura 147, se distingue el barrio de Tiro de Línea al sureste, debido al color amarillo de los puntos que identifican la posición de los inmuebles.

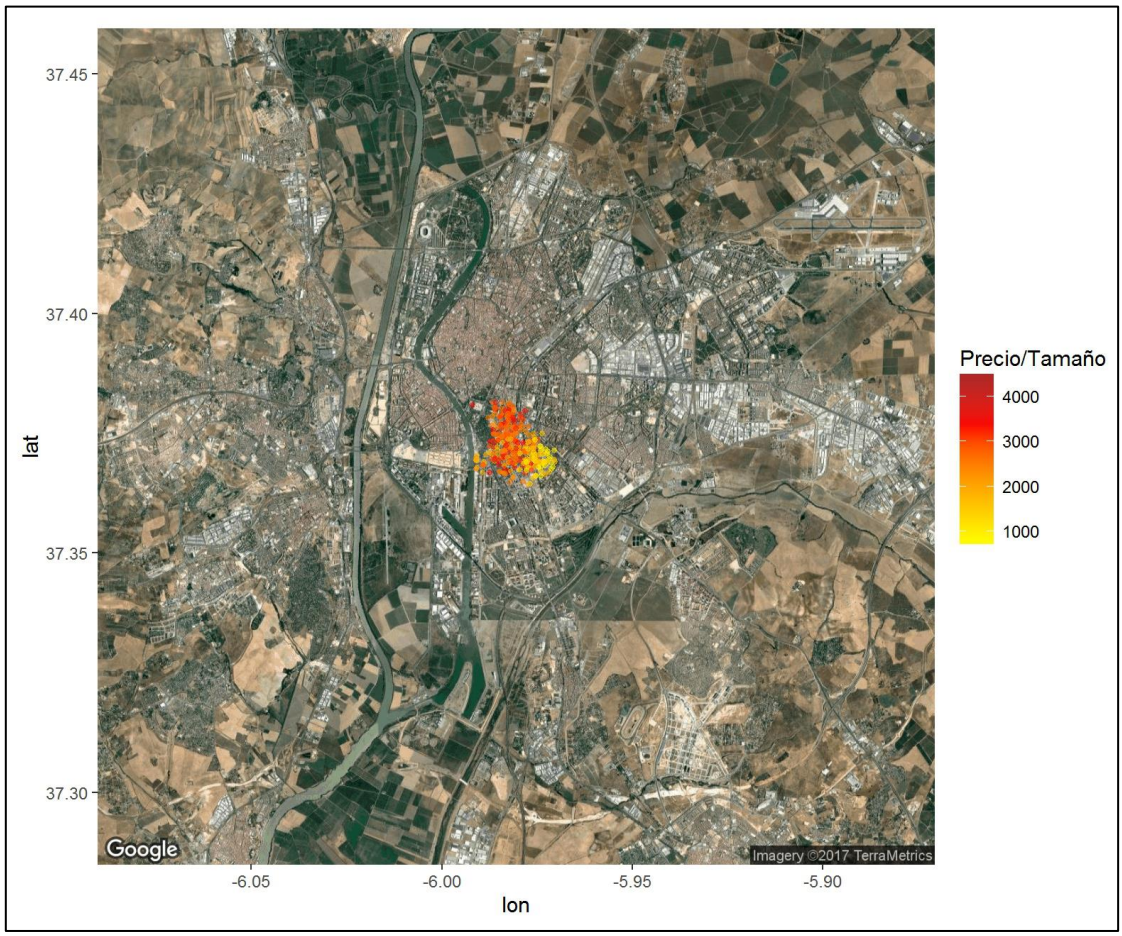


Figura 147. Distribución geográfica. Precio por metro cuadrado. Prado de San Sebastián – Felipe II. (Fuente: Elaboración propia)

8.2.2.11. Distrito San Jerónimo

Esta barriada se encuentra en el extremo norte de la ciudad. Al oeste de Pino Montano, y perteneciente junto a este, al distrito Norte de la ciudad. Está situada en la orilla oriental del río Guadalquivir a su paso por la ciudad.

Se ha recogido información de un total de 81 viviendas que se encuentran a la venta en esta zona, para las que se ha calculado un precio medio por metro cuadrado de 1.009,97 €, por lo que es la tercera zona con el precio medio más bajo de la ciudad, tras Cerro Amate y Torreblanca.

Las viviendas analizadas en esta zona se muestran en la Figura 148.

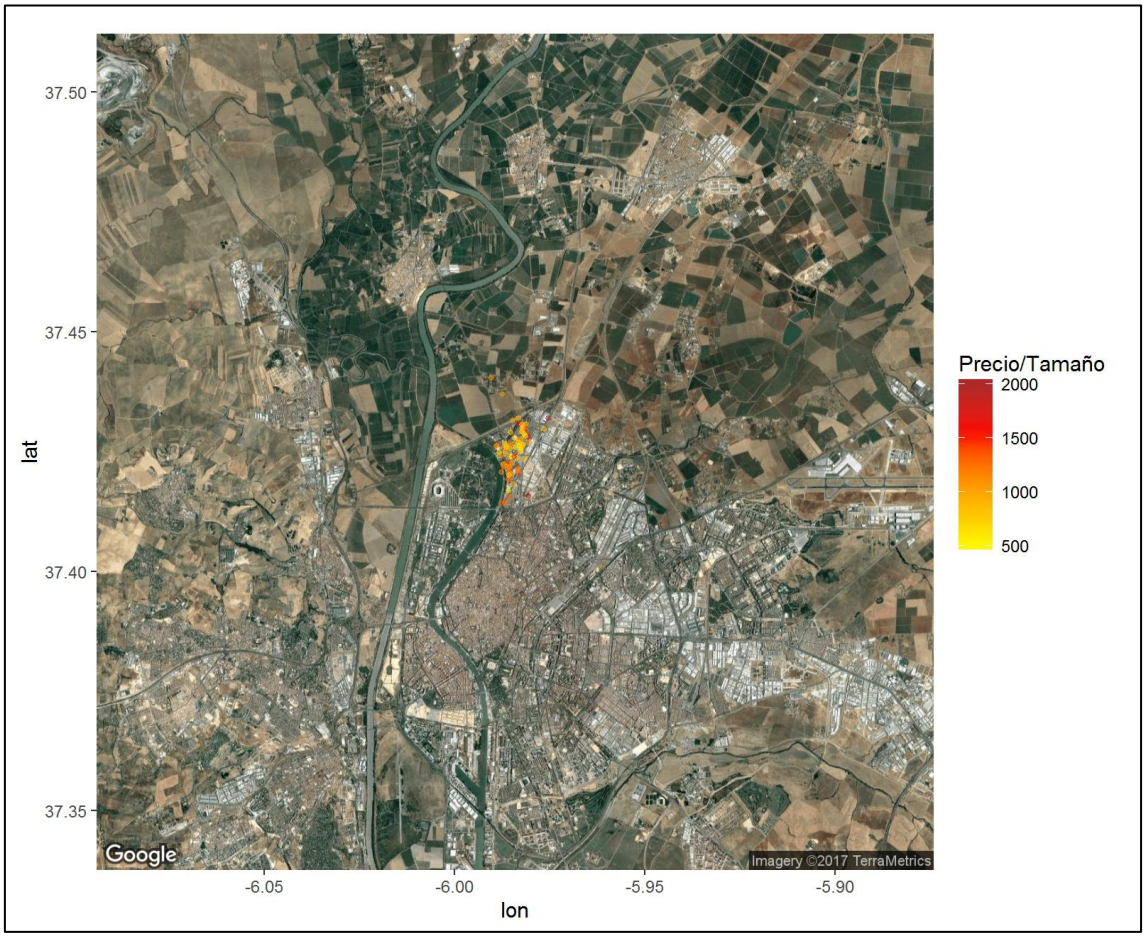


Figura 148. Distribución geográfica. Precio por metro cuadrado. San Jerónimo. (Fuente: Elaboración propia)

8.2.2.12. Distrito San Pablo

El conocido como Polígono de San Pablo, pertenece al distrito administrativo de San Pablo – Santa Justa, que en este análisis se ha dividido, analizando la zona de Santa Justa posteriormente.

Se encuentra situado en la zona este de la ciudad, al nordeste del distrito de Nervión. Se han analizado los precios de 213 viviendas de esta zona, arrojando un precio medio de 1.145,94 € / m^2, es decir, un valor significativamente inferior al de la ciudad y prácticamente igual que el calculado en la zona de Pino Montano.

Como puede observarse en la Figura 149, el precio de los inmuebles de la zona es muy homogéneo. Sólo destacan por tener un precio más alto, algunas viviendas de la zona norte de la barriada.

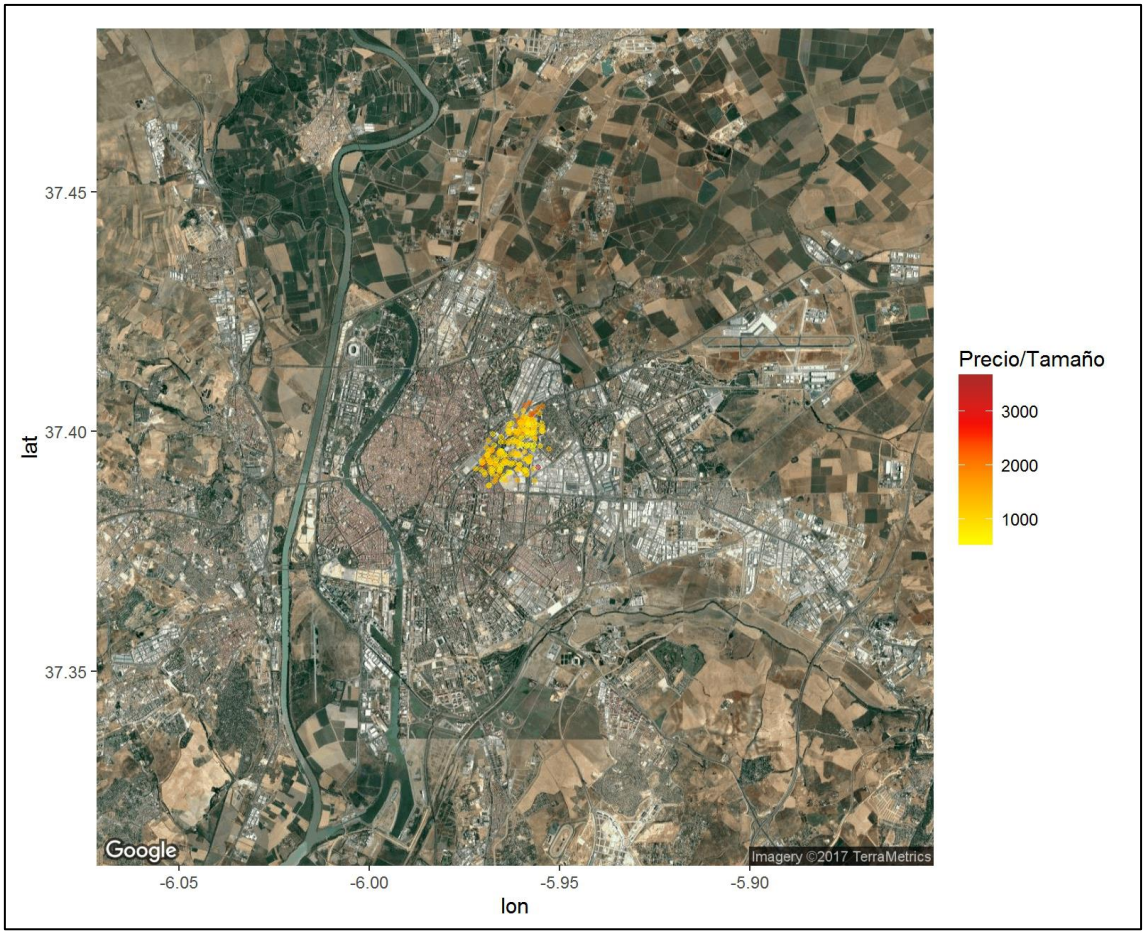


Figura 149. Distribución geográfica. Precio por metro cuadrado. San Pablo. (Fuente: Elaboración propia)

8.2.2.13. Distrito Santa Clara

Al este del polígono de San Pablo, se encuentra la barriada de Santa Clara, perteneciente también al mismo distrito que éste. Es un barrio residencial con casas, en su mayoría, unifamiliares, como chalés y dúplex, aunque también dispone de bloques de pisos.

Al ser mayoritaria la presencia de viviendas unifamiliares, el precio medio de la vivienda es superior al encontrado en la barriada de San Pablo, colindante con ésta. Su precio medio asciende a 1.964,12 € / m^2, para una muestra de 112 viviendas.

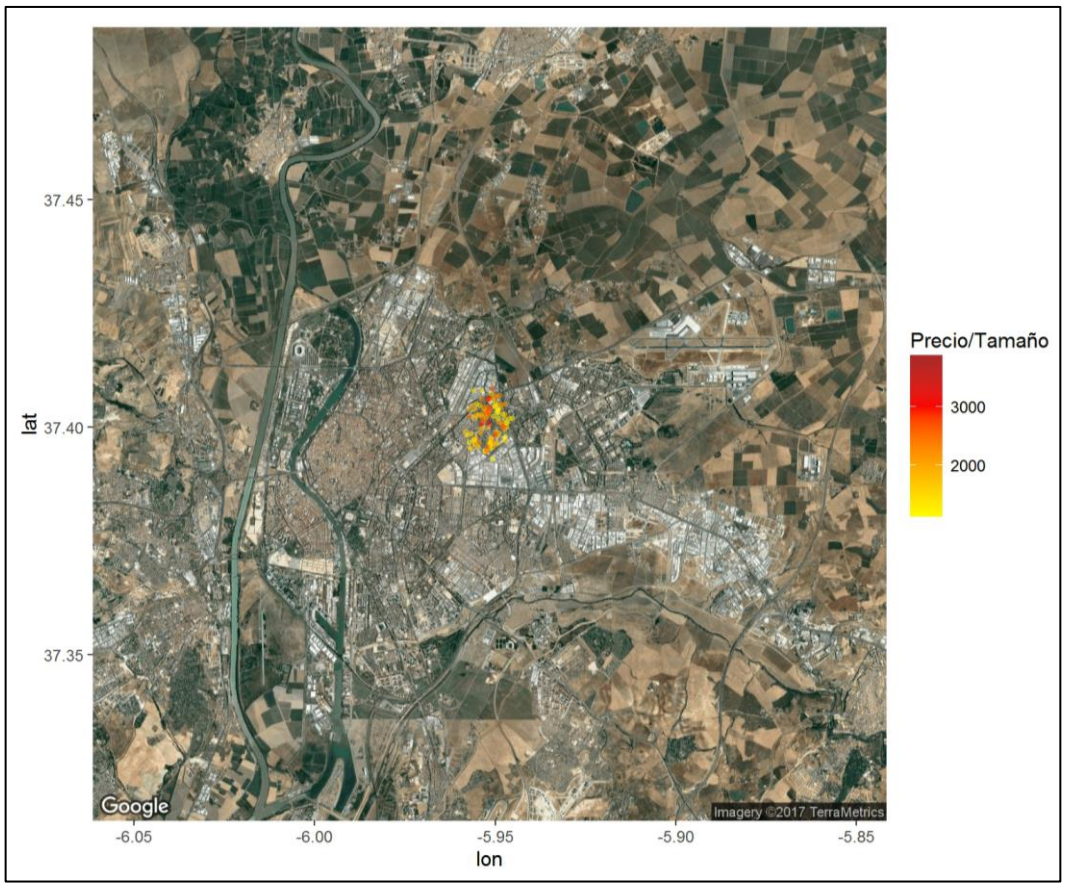


Figura 150. Distribución geográfica. Precio por metro cuadrado. Santa Clara. (Fuente: Elaboración propia)

Aunque este precio medio está situado en valores cercanos a la media, cabe destacar que la tipología de las viviendas de la zona hace de éste, el barrio en el que el precio medio total de las viviendas es mayor en toda la ciudad, con un valor de 520.785 €. Su distribución, en la barriada, puede observarse en la Figura 150.

8.2.2.14. Distrito Santa Justa – Miraflores – Cruz Roja

Esta zona se encuentra al sur del distrito Macarena, y al este de Centro. Pueden distinguirse en ella, principalmente, cuatro barrios: Arroyo – Santa Justa, Cruz Roja – Capuchinos, Miraflores – Carretera de Carmona y La Salle - Avenida Manuel del Valle – Las Naciones.

Se han recogido datos de 521 viviendas de la zona, véase Figura 151, con un precio medio de 1.728,76 € / m^2, ligeramente inferior al precio medio hallado para el conjunto del municipio.

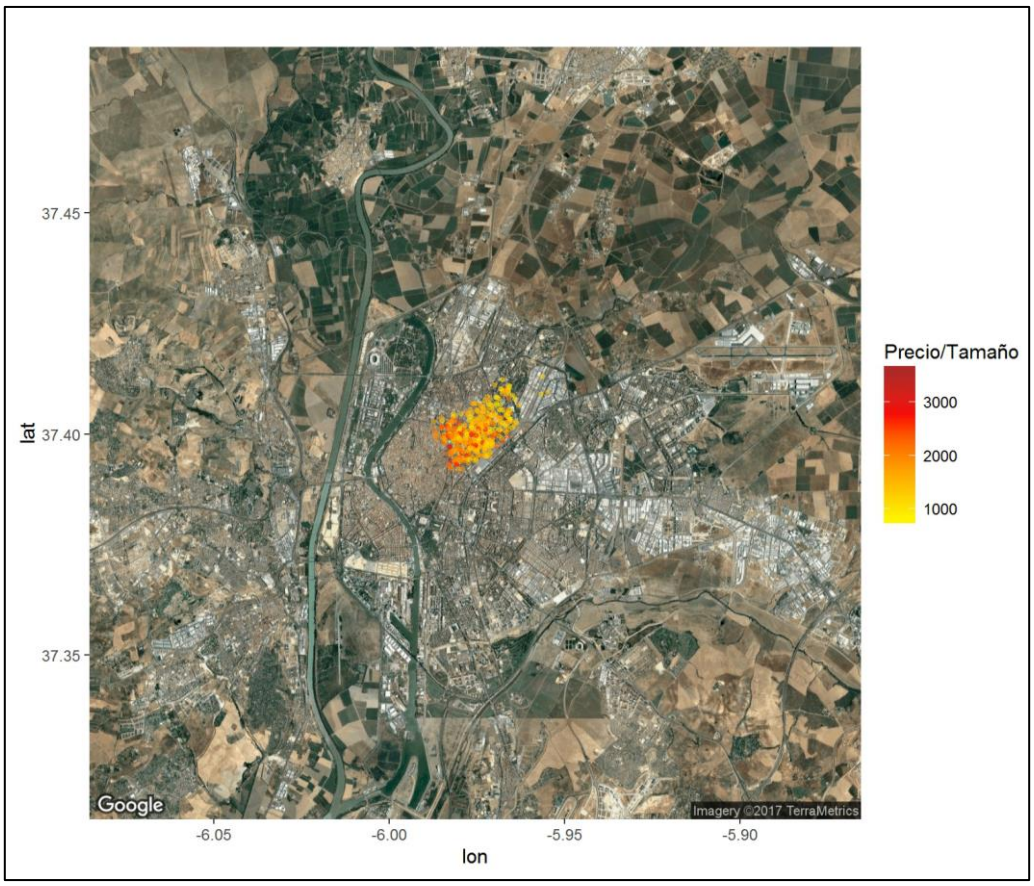


Figura 151. Distribución geográfica. Precio por metro cuadrado. Santa Justa – Miraflores – Cruz Roja. (Fuente: Elaboración propia)

Por barrios, al igual que ha ocurrido en otras zonas, los que limitan con el distrito Centro tienen el precio medio mayor. Este es Cruz Roja – Capuchinos, con 1.916,50 € / m^2, y que limita a su vez con Macarena, y Arroyo – Santa Justa, en la zona sur, pero limitando también por una zona con el distrito Centro, con un precio de 1.777,5 € / m^2.

8.2.2.15. Distrito Sevilla Este

Esta zona, junto con Parque Alcosa y Torreblanca (Véase *Figura 152*), forman el denominado distrito Este. Se encuentra en la zona oriental de la ciudad, al sur de Parque Alcosa.

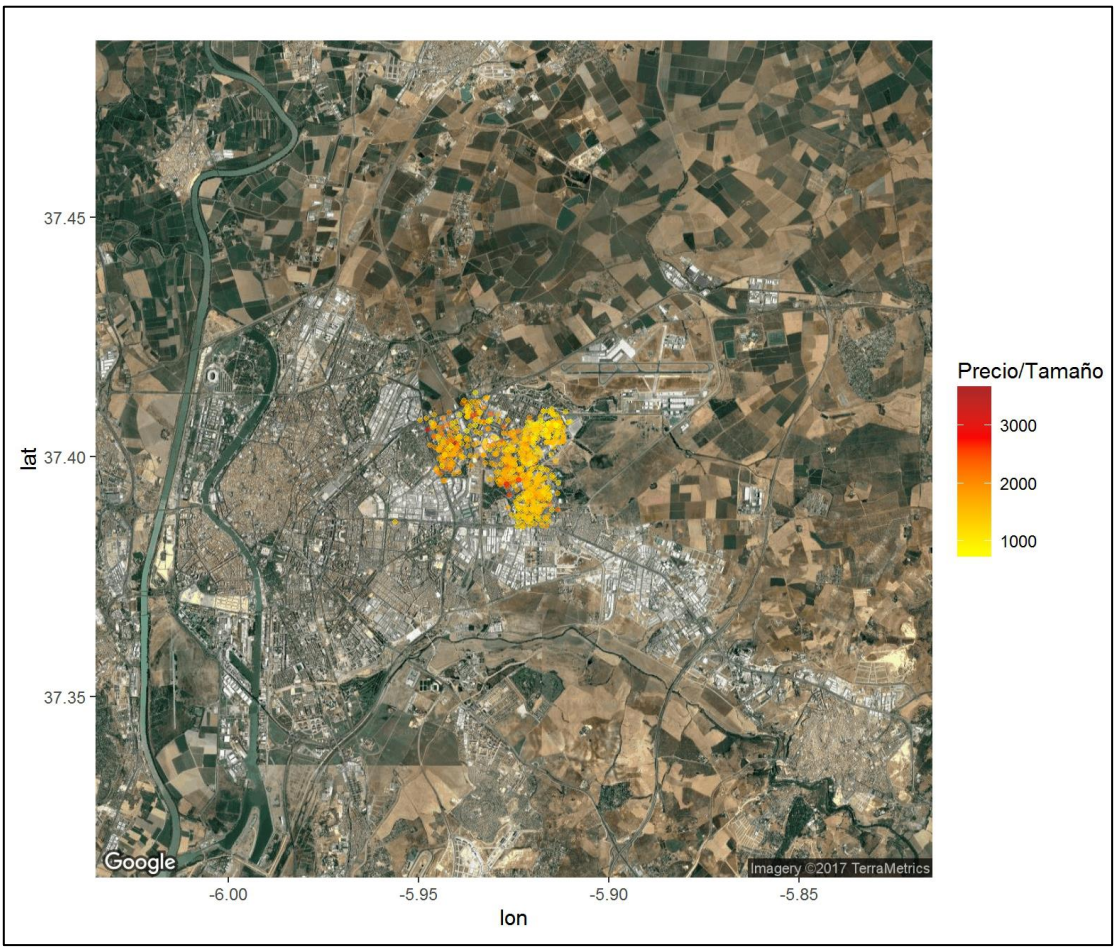


Figura 152. Distribución geográfica. Precio por metro cuadrado. Sevilla Este. (Fuente: Elaboración propia)

Para un total de 717 viviendas analizadas, el precio medio hallado es de 1489,91 € / m^2, precio inferior a la media, de esperar, al ser un barrio situado en la periferia de la ciudad.

Puede dividirse en tres barrios: Alcalde Luis Uruñuela – Palacio de Congresos, Avenida de las Ciencias y Emilio Lemos; en la zona norte, sur y este, respectivamente.

En la zona del palacio de Congresos y la avenida Alcalde Luis Uruñuela, el precio medio es significativamente superior a las otras, entre las que no se encuentran

diferencias. No obstante, como ocurre en Sevilla Este, el precio medio en las tres zonas es inferior a la media, estando por debajo de los 1.650 € en la más cara, y en torno a los 1.400 € en las otras dos.

8.2.2.16. Distrito Torreblanca

El barrio de Torreblanca es una zona del extrarradio, situada al este de la ciudad, dividido por la autovía A92, y rodeado de polígonos industriales. Limita al este con Cerro Amate y Sevilla Este. Se trata de un barrio deprimido de la ciudad, con problemas socioeconómicos, agravados por la crisis económica reciente. No en vano, presenta el precio medio por metro cuadrado más bajo de ésta, junto con el barrio de Los Pajaritos en Cerro Amate.

Sobre 85 inmuebles analizados, el precio medio es de 706,15 € / m^2, casi un tercio del precio medio de la ciudad.

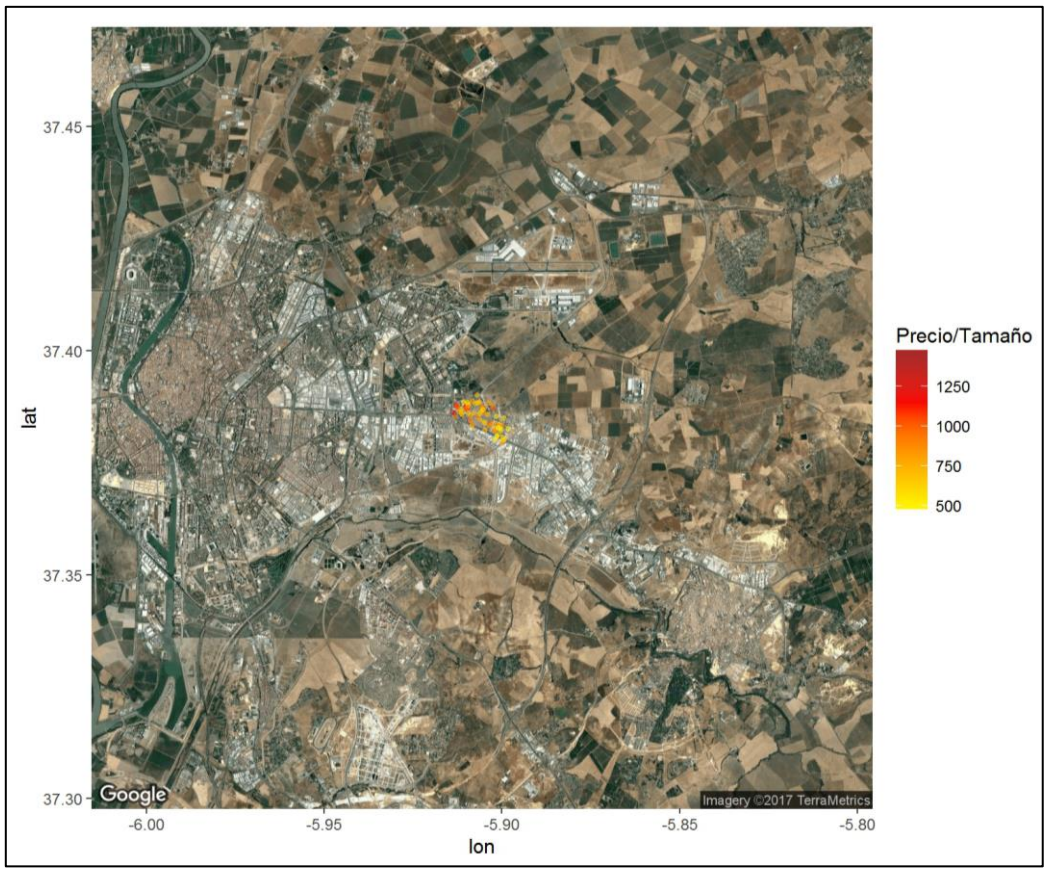


Figura 153. Distribución geográfica. Precio por metro cuadrado. Torreblanca. (Fuente: Elaboración propia)

En él puede encontrarse, por ejemplo, un piso de 66 metros cuadrados por un precio de 20.000 €, lo que supone un precio por metro cuadrado de 303 €.

La distribución de viviendas en el barrio y la situación de éste en la ciudad puede verse en la Figura 153.

8.2.2.17. Distrito Triana

Es el distrito más conocido y emblemático de la ciudad de Sevilla (*Figura 154*). Está situado en la orilla occidental del río Guadalquivir, al norte del distrito de Los Remedios.

Se han hallado un total de 640 viviendas a la venta en el distrito de Triana. Su precio medio por metro cuadrado es de 2.185,95 €.

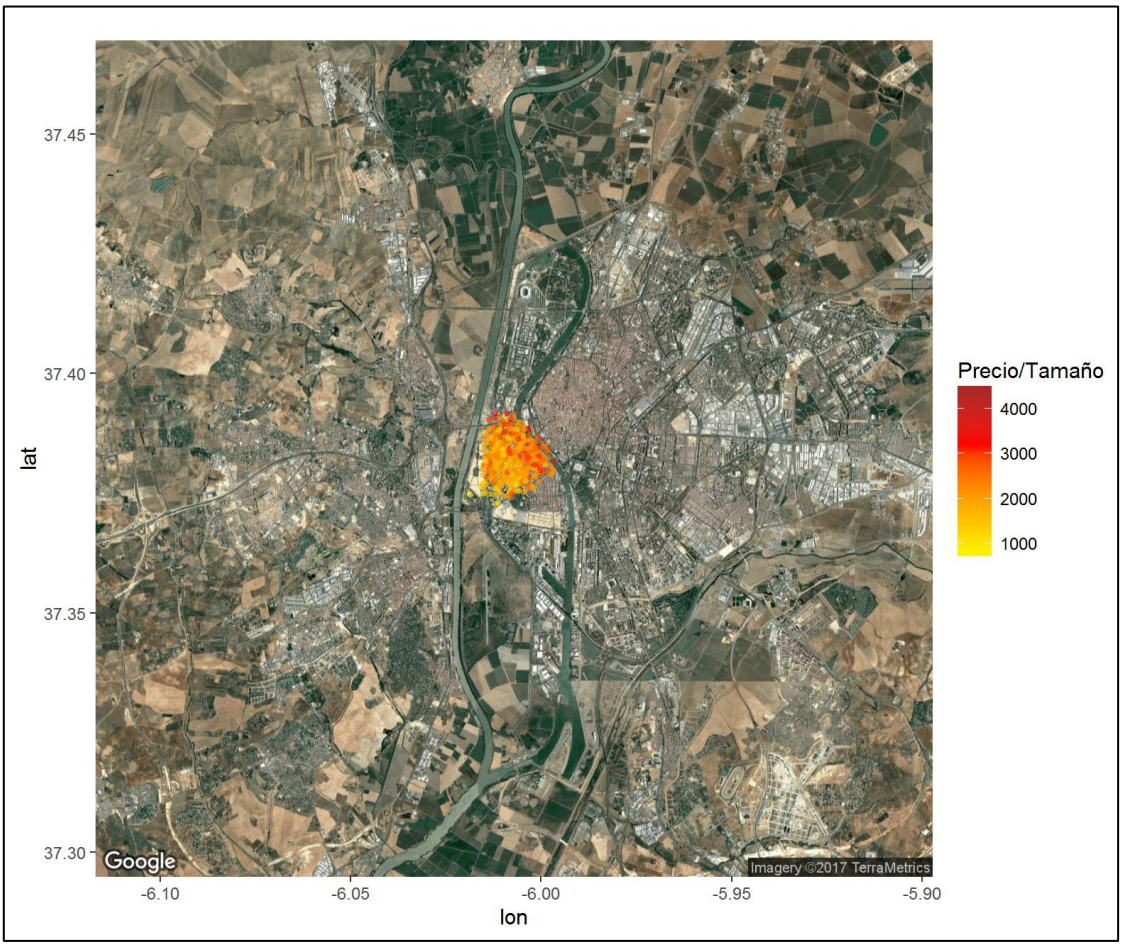


Figura 154. Distribución geográfica. Precio por metro cuadrado. Triana. (Fuente: Elaboración propia)

Este distrito puede dividirse, a su vez, en cuatro barrios o zonas diferenciadas: Altozano – Pagés del Corro, situado a la orilla del río; Barrio León – Tardón, al suroeste; Pagés del Corro – López de Gomara, al sudeste y Ronda de Triana – Patrocinio – Turruñuelo, al noroeste.

La zona a orillas del río Guadalquivir tiene un precio medio considerablemente superior al resto, situándose éste en 2.517,78 € / m^2. El resto de zonas se mantiene en precios medios por encima de los 2.000 €, salvo Barrio León – Tardón, que tiene un precio medio de 1.920,83 €.

8.2.3. Otras características

A continuación, analizaremos el resto de características de los inmuebles sobre las que se dispone de información.

Al analizar el número de habitaciones de las viviendas ofertadas, se obtuvo un valor medio de 3,25 y una mediana de 3. En este caso la dispersión es menor y el 90 % de las viviendas disponían de un número de habitaciones comprendido entre 1 y 6. El intervalo de confianza al 95 % nos devolvió un número medio comprendido entre 3,22 y 3,28 habitaciones.

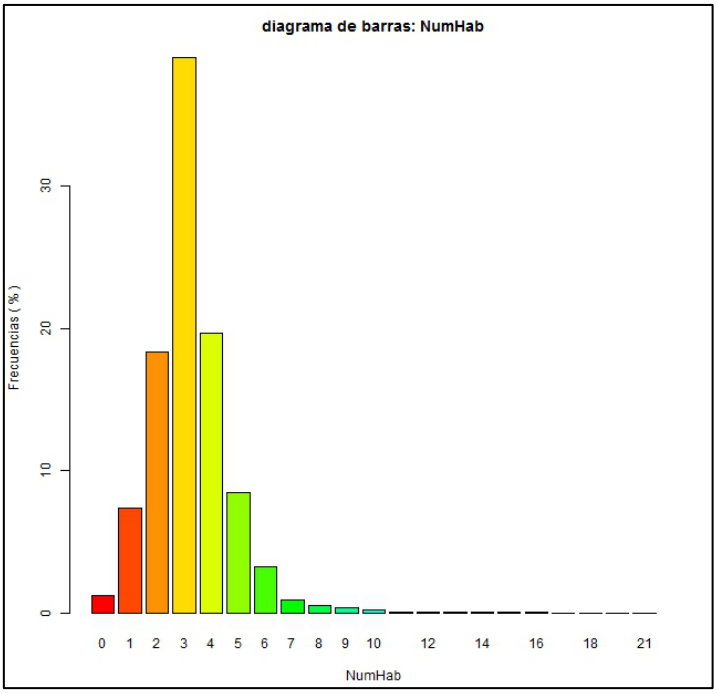


Figura 155. Diagrama de barras. Número de habitaciones. Sevilla. (Fuente: Elaboración propia)

La Figura 155, muestra el diagrama de barras con la distribución del porcentaje de viviendas según el número de habitaciones.

El número de habitaciones de una vivienda suele estar estrechamente relacionado con el número de baños. En la muestra analizada, esto también se cumple como se comprueba al estudiar la relación lineal entre ambas variables. El coeficiente de correlación lineal de Pearson devuelve un valor de 0,6923, lo que supone una relación lineal directa bastante fuerte. El contraste de hipótesis asociado devuelve con una probabilidad límite menor de 10^{-4} que la relación es significativa.

Veamos ahora, en la Tabla 31 el resumen del estudio descriptivo obtenido para el número de baños de un inmueble.

Tabla 31. Estudio descriptivo. Número de baños. Sevilla.

Valores	NumBaños
Media	1,8219
Mediana	2,0000
Cuasidesviación típica	0,9638
Cuasivarianza	0,9289
Rango	12,0000
Coeficiente de variación	0,5290
Coeficiente de asimetría	1,9609
Coeficiente de apuntamiento	8,0539
Ext. inf. IC media 95 %	1,8031
Ext. sup. IC media 95 %	1,8407
Percentil 5	1,0000
Percentil 10	1,0000
Percentil 25	1,0000
Percentil 75	2,0000
Percentil 90	3,0000
Percentil 95	4,0000

Fuente: Elaboración propia

El número medio de baños de las viviendas de la ciudad está comprendido entre 1,8 y 1,84 y, aunque hay en la muestra viviendas con hasta 12 baños, el 95 % de ellas tienen a lo sumo 4. El 83,6 % del total de viviendas tienen entre 1 y 2 baños y, algo más del 10 % tiene 3 baños.

De seis tipos de viviendas analizadas, la mayor parte de ellas, el 78,22 %, son pisos. El siguiente tipo de vivienda, según número, es el formado por los chalés que representan el 13,70 % del total. La Figura 156, muestra la distribución del tipo de vivienda en un diagrama de sectores.

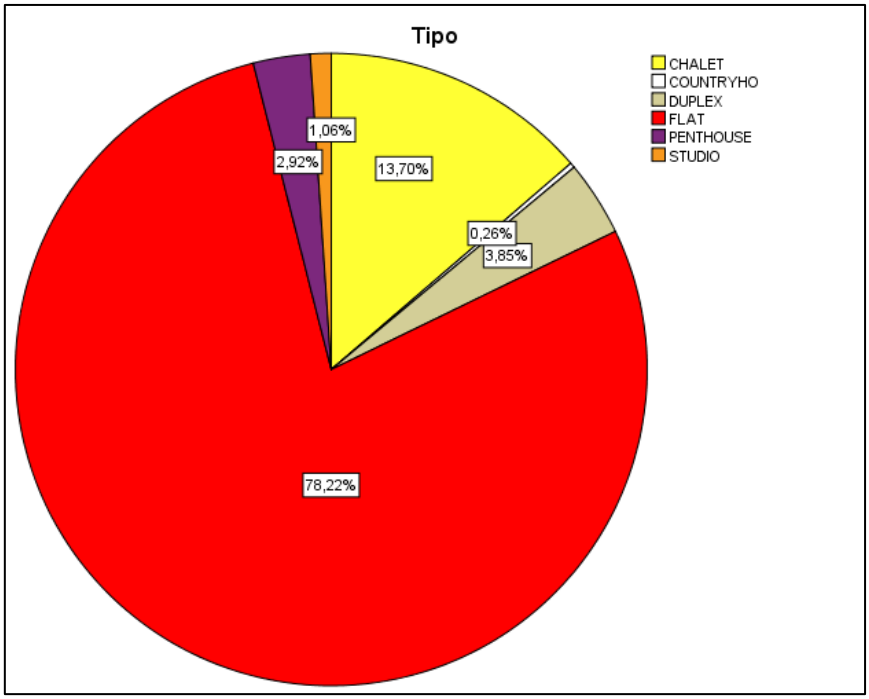


Figura 156. Gráfico de sectores. Tipo de vivienda. (Fuente: Elaboración propia)

La distribución en la ciudad de los distintos tipos de viviendas a la venta, se muestra en la Figura 157, donde, el código de color es el mismo que el utilizado en la Figura 156.

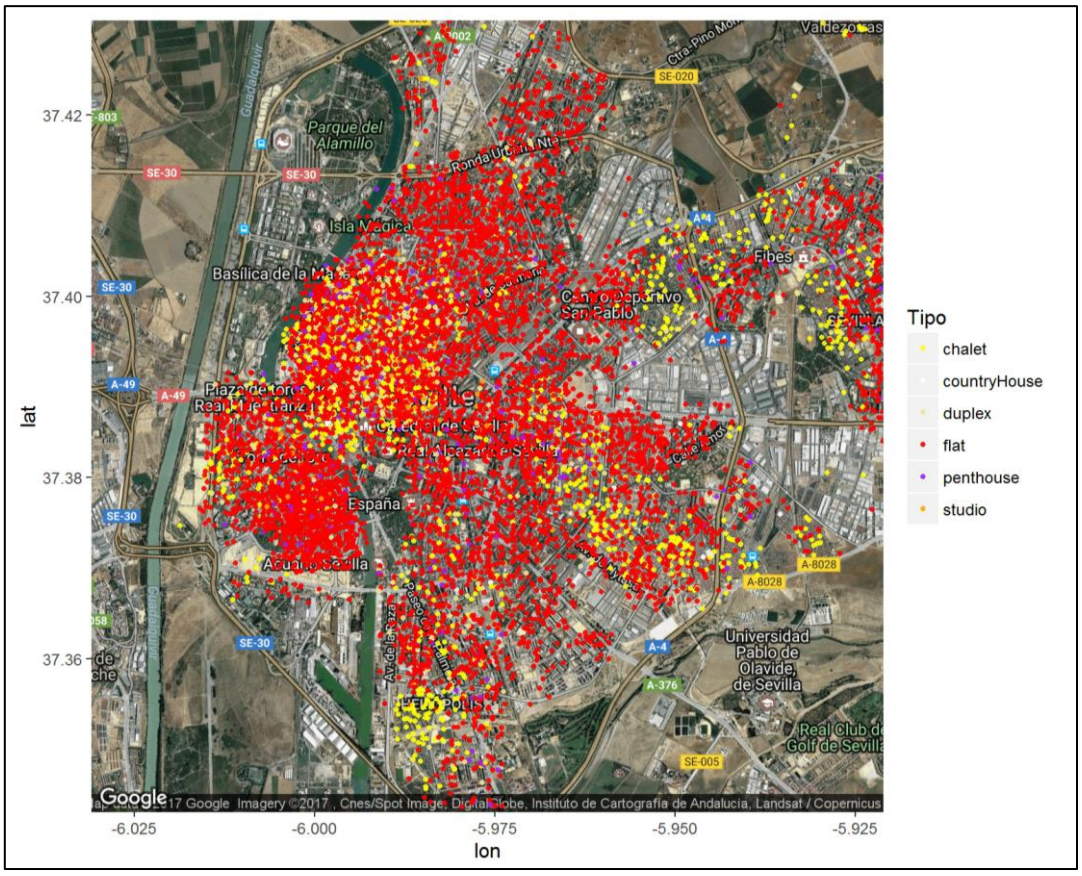


Figura 157. Distribución geográfica. Tipo de vivienda. Sevilla. (Fuente: Elaboración propia)

Los chalés se distribuyen por la ciudad, pero se concentran fundamentalmente en la zona de Heliópolis, en Los Bermejales y en el barrio de Santa Clara y en la zona nororiental de Sevilla Este. Los áticos se concentran fundamentalmente en el distrito de Triana y en la zona centro, aunque también podemos observar su presencia en el barrio de Nervión.

El porcentaje de viviendas que disponen de cochera, ascensor, piscina, aire acondicionado, terraza, trastero y armarios empotrados, y el estado en el que se encuentra, se muestra en la Figura 158.

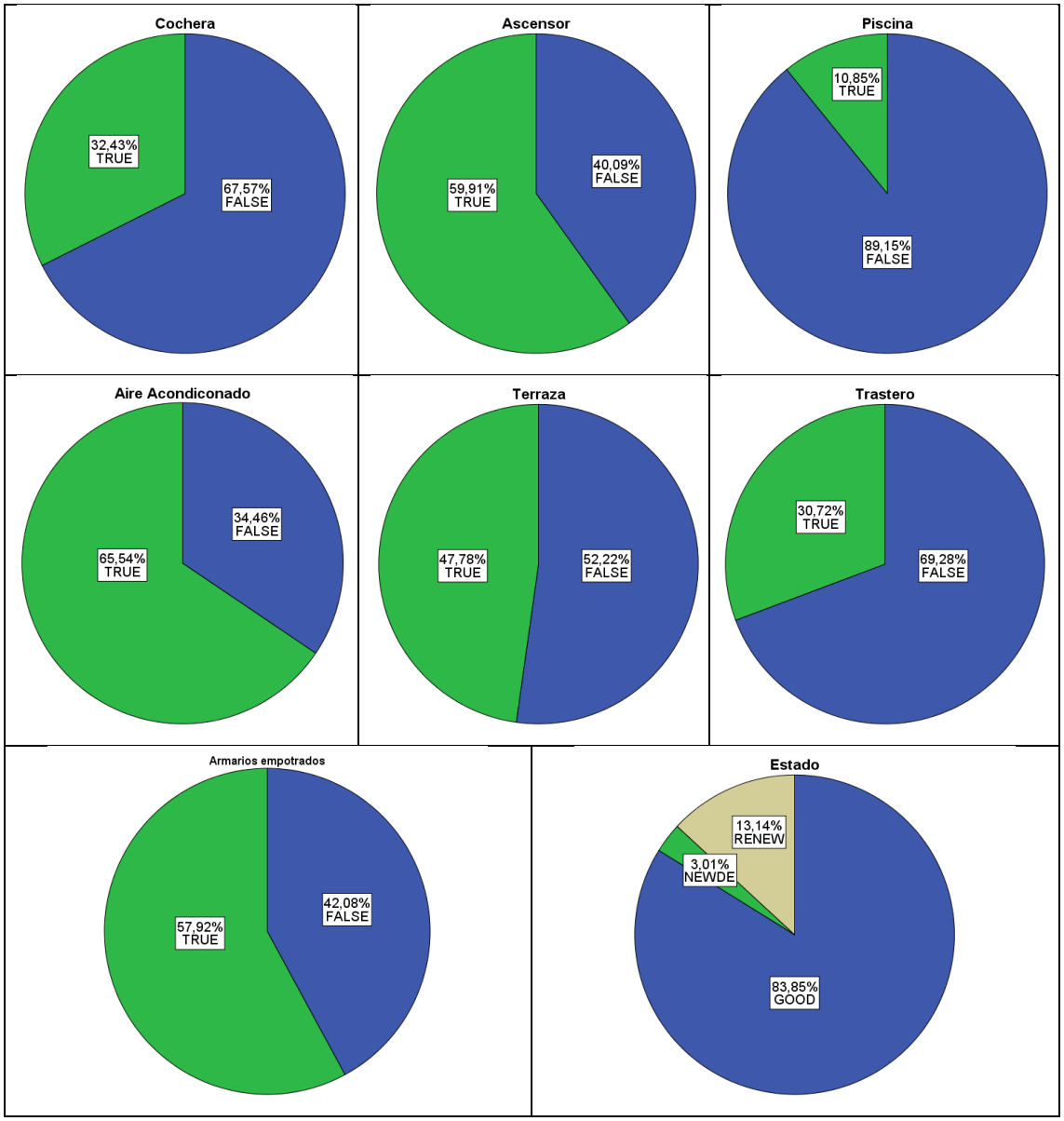


Figura 158.Diagramas de sectores. Características varias. Sevilla. (Fuente: Elaboración propia)

Aproximadamente, un tercio de las viviendas a la venta disponen de cochera, de gran valor en algunas zonas, debido a la escasez de aparcamiento público disponible.

El 59,91 % de las viviendas se encuentran situadas en edificios con ascensor. Este porcentaje aumenta hasta el 69,57 % entre los pisos, estudios y áticos.

Sólo el 10,85 % de las viviendas disponen de piscina. Este porcentaje aumenta hasta el 18 % de los áticos, y el 14,2 % en los dúplex.

Las temperaturas extremas de la ciudad en verano hacen necesario el uso de sistemas de climatización. De ahí que el 65,54 % de las viviendas a la venta disponen de éstos. Este es un porcentaje elevado teniendo en cuenta que la operación ofertada es la venta de la vivienda. Los chalés y las casas rurales disponen de climatización, en menor medida que en el resto de tipos, con porcentajes cercanos al 50 %. Por el otro lado, los áticos, por sus temperaturas extremas, y los dúplex, son, con porcentajes superiores al 80 %, los que más disponen de este sistema.

Un porcentaje ligeramente inferior al 50 % de las viviendas, disponen de terraza; y un porcentaje aún inferior disponen de trastero, el 30,72 %. Las viviendas que en mayor porcentaje disponen de terraza son los áticos, con un 85 % de éstos. Por el contrario, y como cabría esperar, los chalés y las casas de campo son las que más disponen de trasteros, por su mayor tamaño, aproximadamente un 50 % de ellas.

Más de la mitad, casi un 58 % de las viviendas tienen armarios empotrados y de éstas, los dúplex y los áticos, en mayor medida, con un 78,1 % y un 75,5 %.

En relación al estado de la vivienda, una amplia mayoría, el 83,8 % están en buen estado, frente a un 13,1 % que necesita reformas y sólo un 3 % que son de nueva construcción.

Analizando el anuncio de venta, se utiliza, por término medio, 16 fotografías para presentar la vivienda, aunque hay 369, un 3,7 %, que no disponen de fotografías. Los anuncios de chalés, dúplex y áticos son los que más fotografías de la vivienda muestran. Por el contrario, las casas rurales y los estudios son los inmuebles de los que menos fotografías se muestran. La Tabla 32, resume la información relativa al número de fotografías de las viviendas a la venta ofertadas.

Tabla 32. Estudio descriptivo. Número de fotografías. Sevilla.

Valores	NumFotos
Media	16,0193
Mediana	14,0000
Cuasidesviación típica	10,3363
Cuasivarianza	106,8386
Rango	58,0000
Coeficiente de variación	0,6452
Coeficiente de asimetría	0,6831
Coeficiente de apuntamiento	-0,1868
Ext. inf. IC media 95 %	15,8175
Ext. sup. IC media 95 %	16,2211
Percentil 5	1,0000
Percentil 10	4,0000
Percentil 25	9,0000
Percentil 75	22,0000
Percentil 90	32,0000
Percentil 95	38,0000

Fuente: Elaboración propia

Veamos a continuación, el grado de asociación entre diferentes variables cualitativas analizadas. Para ello, se han calculado las tablas de doble entrada de las variables a analizar, para éstas se ha aplicado el test Chi – cuadrado de Pearson para analizar la existencia de relación, y por último se ha decidido el sentido de ésta, a partir del estudio de los residuos estandarizados corregidos.

Se ha hallado asociación entre el distrito al que pertenece una vivienda y su tipo, de forma que en el distrito Centro la proporción de dúplex es superior al del resto y el de pisos inferior, mientras que en Santa Clara predominan los chalés frente al resto de zonas.

Por el contrario, en Los Remedios y Macarena hay un predominio de los pisos, por encima del resto de distritos. Respecto a los áticos y los estudios, no se han hallado grandes diferencias en las proporciones encontradas.

Al analizar la disponibilidad de plaza de garaje o cochera, observamos que, en Sevilla Este, y, en menor medida, en Santa Clara, Prado de San Sebastián – Felipe II y La Palmera – Los Bermejales, la proporción de viviendas que disponen de ella es superior. Macarena, Centro y Triana tienen una proporción de viviendas con cochera inferior al global, coincidiendo con una mayor proporción de pisos.

Respecto a la existencia de piscina, salvo las diferencias mostradas en las zonas en las que hay una mayor proporción de viviendas unifamiliares, destaca Sevilla Este, con un gran número de edificios de viviendas con piscina comunitaria.

Por otro lado, el porcentaje de viviendas con terraza en el distrito Centro, es significativamente inferior y, en el Los Remedios, superior. En este último distrito, es donde, además, hay una mayor proporción de viviendas a la venta que necesitan reforma. El distrito Centro, por el contrario, es en el que mayor proporción están en buen estado. Sevilla Este, es el distrito en el que mayor es el número de viviendas de nueva construcción hay en oferta, en proporción a las disponibles.

Es curioso observar que, al comparar el tipo de vivienda con la disponibilidad de aire acondicionado en la misma, el resultado nos indica que la proporción de chalés que disponen de él es inferior, mientras que el de dúplex es superior al resto.

Algo similar ocurre cuando comparamos el tipo de vivienda con el estado de la misma. Mientras que la proporción de chalés que necesitan ser reformados es elevada, la de dúplex es significativamente inferior al resto.

Por último, cabe decir que, al comparar la disponibilidad de cochera, ascensor, piscina, aire acondicionado, terraza, trastero y armarios empotrados, dos a dos; en todos los casos, salvo en el que se comparaba ascensor y trastero, en el que el resultado no es concluyente, podemos afirmar que la proporción de inmuebles que disponen de ambas características o de ninguna de ellas, es superior al porcentaje que dispone únicamente de una de las características.

8.2.4. Estimación del precio de la vivienda de la ciudad de Sevilla

El objetivo ahora, es determinar un modelo que nos permita estimar el precio de oferta de una vivienda de la ciudad de Sevilla en función de las características de que dispone, de forma que consideremos el valor final del bien como la suma de las valoraciones de los distintos elementos que lo caracterizan. Por tanto, aplicaremos la metodología de los precios hedónicos para su estimación.

Se ha eliminado de este estudio, por sus características diferenciadoras y por su escasa representación en la muestra, las casas rurales o de campo, ya que sólo se disponen de 25 inmuebles de este tipo.

En primer lugar, crearemos un modelo lineal general que explique el precio de oferta de todos los tipos de viviendas, mediante un modelo lineal general.

Tras numerosas pruebas realizadas, el mejor modelo hallado, explica el precio de la vivienda en función de tres variables cuantitativas que son Distancia al centro geográfico de la ciudad, Número de baños y Tamaño del inmueble; y siete variables cualitativas: Estado de la vivienda, y las relativas a la existencia de cochera, ascensor, piscina, aire acondicionado, armario empotrado y trastero.

Para la inclusión de éstas últimas variables se han generado variables dicotómicas que permitan su estudio. De esta forma, las variables auxiliares incluidas en el modelo han sido:

- CocheraFALSE: Toma el valor 1 si la vivienda carece de cochera.
- AscensorFALSE: Toma el valor 1 si la vivienda no dispone de ascensor.
- PiscinaFALSE: Toma el valor 1 si la vivienda carece de piscina.
- AireaconFALSE: Toma el valor 1 si la vivienda carece de piscina.
- EmpotradosFALSE: Toma el valor 1 si la vivienda carece de piscina.
- TrasteroFALSE: Toma el valor 1 si la vivienda no dispone de trastero.
- Para el estado de la vivienda:
    - Estadogood: Toma el valor 1 si la vivienda es de segunda mano, pero está en buen estado.
    - Estadonewdevelopment: Toma el valor 1 si la vivienda es de nueva construcción.

Por tanto, si la vivienda necesita reforma, estas dos últimas variables tomarán el valor 0.

Podemos afirmar que el modelo construido es globalmente válido, tras el cálculo del coeficiente de determinación y el contraste ANOVA global, cuyos resultados se muestran en la Tabla 33.

Como puede observarse, el modelo construido estima el 72,15 % de la variabilidad en el precio de oferta de una vivienda de la ciudad de Sevilla. El valor del criterio de información de Akaike obtenido es igual a 256.899,0553, valor muy elevado debido al gran tamaño de la muestra de estudio, formada por 9.625 inmuebles, algo inferior a la muestra inicial, formada por 10.056, debido a la existencia de valores perdidos en la variable Estado.

Tabla 33. Medidas de ajuste del modelo I. Viviendas. Sevilla.

	Valores
Coeficiente de determinación	0,7215
Coeficiente de determinación ajustado	0,7212
Estadístico F	2263,6860
Grados de libertad núm.	11,0000
Grados de libertad den.	9613,0000
Probabilidad límite	0,0000
Criterio de información AIC	256899,0553

Fuente: Elaboración propia

La limitación de la capacidad de ajuste del modelo, se debe a la presencia valores atípicos y/o observaciones influyentes, que el modelo no puede ajustar y que provoca la desviación del hiperplano de ajuste a la nube de puntos.

Es por ello, que, utilizando el método de detección de observaciones influyentes propuesto en este trabajo, y antes de continuar con la validación del modelo construido, se procederá a identificar dichas viviendas, para decidir, posteriormente, sobre la pertinencia de su eliminación del estudio.

Se ha aplicado el método, definiendo como observaciones influyentes a aquellas que verifican que el valor del criterio de información del modelo obtenido al eliminar dicha observación se desvía inferiormente de la media de todos los AIC(-i), más de tres veces su desviación típica, es decir, todas aquellas observaciones para las que AIC(-i) es inferior a 256.834,347.

Los resultados obtenidos se muestran en la Tabla 34. De esta tabla se ha eliminado la primera columna que contiene los códigos que identifican los inmuebles en el portal inmobiliario, pero que nos permite identificarlos con facilidad, de modo que podamos realizar el filtrado de la muestra, en caso de que decidamos eliminarlos.

Como se puede observar, todos los inmuebles detectados como influyentes en el modelo, son poco representativos del conjunto, debido a que, de las 34 viviendas seleccionadas, 32 tienen precios muy elevados, a partir del millón y medio de euros. Las dos viviendas restantes, con precios de 245.000 € y 495.000 €, tienen un tamaño muy elevado para su precio, con 1.125 m^2 y 2.220 m^2 respectivamente.

Tabla 34. Observaciones influyentes. Estimación del precio de las viviendas de Sevilla.

AIC(-i)	Precio	Distancia	NumBaños	Tamaño
256821,871	1700000	1783	3	270
256791,989	2000000	3455	3	300
256749,416	2700000	3001	4	524
256828,734	2080000	1788	5	530
256816,022	2000000	2202	4	405
256829,457	2250000	622	6	530
256744,633	2500000	1242	2	425
256827,575	2300000	1456	6	611
256764,801	3000000	1860	3	800
256799,662	2640000	2470	4	680
256755,71	245000	2483	2	1125
256272,184	2500000	1364	20	3000
256823,567	2300000	1177	5	578
256832,495	1995000	1479	3	528
256818,663	1500000	1254	2	170
256827,821	2000000	1125	3	484
256819,076	2100000	638	4	450
256818,736	1700000	1981	3	250
256797,055	2250000	745	4	425
256405,01	495000	5076	3	2220
256782,483	2575000	789	0	748
256804,935	1700000	1906	10	1500
256795,413	3100000	741	1	1065
256766,955	3000000	2392	5	700
256818,736	1700000	1981	3	250
256748,592	2780000	1622	4	524
256814,077	2800000	596	10	608
256714,385	3045000	2376	7	450
256795,103	2300000	573	6	350
256339,828	7100000	1261	10	1980
256316,27	1900000	3261	6	3100
256827,288	2200000	2218	2	700
256827,217	2600000	679	8	635
256788,173	2200000	3291	4	370

Fuente: Elaboración propia

Aunque pudiera parecer que el método se ha limitado a identificar las viviendas de mayor precio, observando la muestra, hay un total de 98 viviendas con precio igual o superior 1.500.000 €, de las que sólo 32 han sido consideradas atípicas.

Al analizar con detenimiento los inmuebles no identificados como influyentes, se observa que no se han considerado como tales por motivos como que al alto precio le asocia un gran tamaño y un elevado número de baños. La Tabla 35, muestra las viviendas con mayor precio que no han sido consideradas como observaciones influyentes.

Tabla 35. Observaciones NO influyentes. Estimación del precio. Viviendas. Sevilla.

Precio	Distancia	NumBaños	Tamaño
3.500.000	2.785	5	1.700
3.100.000	741	8	1.065
2.798.000	833	15	610
2.675.000	1.354	3	700
2.500.000	873	6	750
2.500.000	983	8	700
2.400.000	887	6	792
2.300.000	1.359	9	819
2.300.000	1.264	3	989
2.200.000	3.633	6	700

Fuente: Elaboración propia

La primera de las viviendas, tiene el segundo precio más elevado de la muestra, sin embargo, no es detectada como influyente, debido a que éste es acompañado de un tamaño de 1.700 m^2, que puede, de algún modo, justificarlo. Algo similar ocurre con la vivienda de 3.100.000, ya que posee 8 cuartos de baño y un tamaño de 1.065 m^2. La tercera vivienda de la lista, posee 15 cuartos de baño. Las que tienen un precio de 2.500.000 € tampoco son identificadas como influyentes debido, fundamentalmente a que se acompaña de un elevado tamaño y número de baños.

Se eliminan las 34 observaciones detectadas como influyentes y se vuelve a estimar el modelo de regresión, ahora con 9.591 viviendas. Una vez eliminadas, se ha prescindido de la variable Trastero por ser irrelevante.

Los resultados de las medidas de bondad de ajuste se muestran en la Tabla 36. Este modelo, explica, ahora, el 80,13 % de la variabilidad del precio de oferta de una vivienda del municipio de Sevilla, es decir, ha experimentado un aumentado de casi ocho puntos respecto al modelo anterior, al eliminar un 0,35 % de las observaciones.

Tabla 36. Medidas de ajuste del modelo II. Viviendas. Sevilla.

	Valores
Coeficiente de determinación	0,8013
Coeficiente de determinación ajustado	0,8011
Estadístico F	3864,2859
Grados de libertad núm.	10,0000
Grados de libertad den.	9580,0000
Probabilidad límite	0,0000
Criterio de información AIC	250288,5414

Fuente: Elaboración propia

Además, el valor del criterio de información ha disminuido un 2,57 % tras la criba de valores realizada. Los resultados de la estimación de los coeficientes se muestran en la Tabla 37.

Tabla 37. Estimación de coeficientes del modelo II. Viviendas. Sevilla.

	Coeficientes	**Error estándar**	**T**	**P-límite**
Constante	46324,4494	6844,9791	6,7677	0,0000
Distancia	-36,3545	0,7900	-46,0187	0,0000
Tamaño	1903,2007	18,8974	100,7123	0,0000
NumBaños	43020,7688	1886,7616	22,8014	0,0000
CocheraFALSE	-16162,5832	2868,8384	-5,6338	0,0000
AscensorFALSE	-21547,1956	2553,6277	-8,4379	0,0000
PiscinaFALSE	-32514,9105	3980,8159	-8,1679	0,0000
AireAconfalse	-9327,9134	2909,9308	-3,2055	0,0014
Estadogood	51785,5659	3617,5368	14,3151	0,0000
Estadonewdevelopment	60805,5497	7576,8218	8,0252	0,0000
EmpotradosFALSE	-11164,7795	2828,9502	-3,9466	0,0001

Fuente: Elaboración propia

Todas las variables son relevantes en el modelo ya que su eliminación perjudicaría la capacidad de ajuste de éste por lo que disminuiría el valor del coeficiente de determinación ajustado y aumentaría el valor del criterio de información.

La inclusión de la información relativa al distrito mejoraría ligeramente la capacidad predictiva del modelo, pero lo haría muy complejo, ya que, al considerar 17 zonas distintas, éste introduciría 16 variables auxiliares en éste. No obstante, las peculiaridades

de cada distrito hacen más recomendable generar modelos de estimación para cada uno de forma independiente.

La capacidad predictiva se ha medido a través del cálculo del error absoluto medio y el error relativo medio, para una muestra de validación de 902 viviendas. El resultado del primero es 64.944,95 €, mientras que el del segundo es 0,3836.

Los contrastes de validación muestran ausencia de multicolinealidad en el modelo. El factor de inflación de la varianza más elevado se ha dado para la variable NumBaños, con un valor de 2,45, y el índice de condición normalizado ha devuelto un valor de 17,78. El test de Goldfeld-Quandt ha relevado ausencia de heterocedasticidad en los residuos. Por el lado negativo, los test de normalidad revelan ausencia de normalidad en los residuos, lo que lleva a pensar en la necesidad de acotar el precio de la vivienda a analizar.

El modelo estimado es el siguiente:

$$\widehat{Precio} = 46324{,}4494 - 36{,}3545 \cdot Distancia + 43020{,}7688 \cdot NumBaños + 1903{,}2007 \cdot Tamaño - 16162{,}5832 \cdot CocheraFALSE - 21547{,}1956 \cdot AscensorFALSE - 32514{,}9105 \cdot PiscinaFALSE - -9327{,}9134 \cdot AireAconfalse - 11164{,}7795 \cdot EmpotradosFALSE + 51785{,}5659 \cdot Estadogood + 60805{,}5497 \cdot Estadonewdevelopment$$

Del mismo modo, se han calculado intervalos de confianza para los coeficientes, cuyos resultados se muestran en la Tabla 38.

Tabla 38. Intervalos de los coeficientes del modelo II. Viviendas. Sevilla

	2.5 %	97.5 %
Constante	32906,8416	59742,0571
Distancia	-37,9031	-34,8060
Tamaño	1866,1578	1940,2436
NumBaños	39322,3166	46719,2209
CocheraFALSE	-21786,1137	-10539,0528
AscensorFALSE	-26552,8464	-16541,5448
PiscinaFALSE	-40318,1522	-24711,6688
AireAconfalse	-15031,9937	-3623,8331
Estadogood	44694,4282	58876,7035
Estadonewdevelopment	45953,3754	75657,7241
EmpotradosFALSE	-16710,1206	-5619,4384

Fuente: Elaboración propia

Analizando estos resultados, podemos afirmar con un 95 % de confianza que, dejando el resto de condiciones de la vivienda invariantes:

- El aumento de la distancia al centro geográfico disminuye el precio, de forma que, el aumento de 1 metro la distancia al centro geográfico, supone una pérdida del valor de la vivienda de entre 34,81 € y 37,90€.
- El aumento de la superficie de la vivienda en 1 metro cuadrado, genera un aumento de su valoración de entre 1.866,16 € y 1.940,24 €, en consonancia con los datos dados al principio de este capítulo sobre el precio de la vivienda en la ciudad de Sevilla.
- Un mayor número de baños, incrementa el precio de la vivienda; que está, a su vez, relacionado con que el aumento en el tamaño de la vivienda. Este aumento supone entre 39.322 € y 46.719 €.
- La ausencia de cochera genera una pérdida de valoración de entre 10.539 € y 21.786 €.
- La ausencia de ascensor supone que la vivienda pierda entre 16.542 € y 26.553 € en su valoración.
- Las viviendas que disponen de piscina, tienen una valoración superior a las que no disponen de ella, de entre 24.712 € y 40.318 €.
- La disponibilidad de climatización en la vivienda supone un incremento en su precio de entre 3.624 € y 15.032 €
- La existencia, en la vivienda, de armarios empotrados aumenta su valoración, entre 5.619 € y 16.710 €.
- Las viviendas en buen estado y las de nueva construcción incrementan su valor al compararlas con las que necesitan reforma. Una vivienda de segunda mano en buen estado tiene un valor superior de entre 44.694 € y 58.877 € que las que necesitan ser reformados, mientras que en las de nueva construcción este incremento asciende a valores comprendidos entre 45.953,38 € y 75.658 €

Para este mismo conjunto de datos, se ha construido una red neuronal con una capa oculta, formada por tres neuronas, con función de activación logística, que ha logrado la convergencia en 29.631 iteraciones, para un umbral de error de 0,01. Los resultados se muestran en la Tabla 39.

A continuación, en la *Figura 159*, podemos observar la arquitectura de la red. Con esta red se ha obtenido un error cuadrático medio inferior, para una muestra de validación del 10 % de la muestra, que el obtenido para el modelo de regresión previamente estimado.

Tabla 39. Red neuronal. Estimación del precio de la vivienda.

Red neuronal	Valores	Red neuronal	Valores
Error	3,5631	**AscensorFALSE.to.1layhid2**	0,1129
reached.threshold	0,0098	**PiscinaFALSE.to.1layhid2**	0,2736
Steps	29631	**AireAconfalse.to.1layhid2**	0,0993
Aic	81,1262	**EmpotradosFALSE.to.1layhid2**	0,1020
Bic	342,4658	**Estadogood.to.1layhid2**	-0,7231
Intercept.to.1layhid1	-9,0299	**Estadonewdevelopment.to.1layhid2**	-0,5724
Distancia.to.1layhid1	2,2416	**Intercept.to.1layhid3**	-0,4716
Tamaño.to.1layhid1	16,5750	**Distancia.to.1layhid3**	3,2922
NumBaños.to.1layhid1	12,9440	**Tamaño.to.1layhid3**	7,1143
CocheraFALSE.to.1layhid1	-6,9731	**NumBaños.to.1layhid3**	0,5283
AscensorFALSE.to.1layhid1	-1,8986	**CocheraFALSE.to.1layhid3**	-0,3647
PiscinaFALSE.to.1layhid1	-0,1598	**AscensorFALSE.to.1layhid3**	0,1591
AireAconfalse.to.1layhid1	0,0325	**PiscinaFALSE.to.1layhid3**	0,3125
EmpotradosFALSE.to.1layhid1	1,5683	**AireAconfalse.to.1layhid3**	0,1248
Estadogood.to.1layhid1	-0,3615	**EmpotradosFALSE.to.1layhid3**	0,1102
Estadonewdevelopment.to.1layhid1	-705,4319	**Estadogood.to.1layhid3**	-0,9922
Intercept.to.1layhid2	0,4285	**Estadonewdevelopment.to.1layhid3**	-0,6508
Distancia.to.1layhid2	2,7986	**Intercept.to.Precio**	0,2840
Tamaño.to.1layhid2	-2,2429	**1layhid.1.to.Precio**	0,2636
NumBaños.to.1layhid2	-0,3293	**1layhid.2.to.Precio**	-0,6971
CocheraFALSE.to.1layhid2	-0,2047	**1layhid.3.to.Precio**	0,4212

Fuente: Elaboración propia

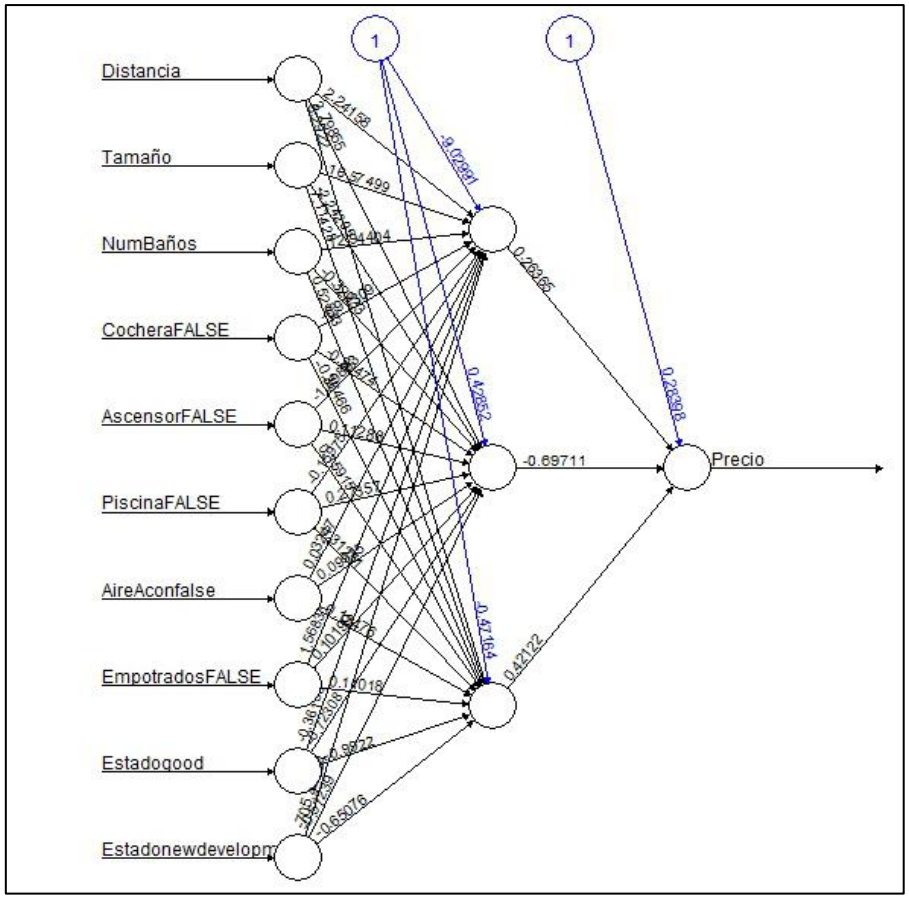


Figura 159. Red neuronal. Arquitectura. Estimación del precio de la vivienda. Sevilla

Para evitar una mayor complejidad, se ha omitido la información relativa a la zona o distrito en la que se encuentra el inmueble. Estimaremos, a continuación, el precio de las viviendas de algunas de las zonas analizadas previamente, mediante un modelo de regresión, detectaremos posibles observaciones influyentes y, por último, construiremos una red neuronal, para realizar una comparación entre ambos modelos.

La selección de zonas está motivada por la disposición de una muestra suficientemente grande, y por la tipología particular de la zona como los tipos de inmuebles que abundan en la misma, o su situación en el centro o en el extrarradio de la ciudad.

8.2.5. Estimación del precio de la vivienda de algunos distritos

A continuación, se estiman se construyen modelos de estimación de precios en cinco distritos con diferentes características socioeconómicas.

8.2.5.1. Bellavista – Jardines de Hércules

Esta zona se encuentra situada en la zona Sur de la ciudad y dispone de un precio medio por debajo del conjunto de la ciudad.

Para una muestra de 248 inmuebles, el modelo obtenido tiene un coeficiente de determinación de 0,7716. Se ha detectado, para éste, 10 observaciones influyentes, que al eliminarlas se obtiene un modelo con un coeficiente de determinación de 0,8261. Para una muestra de validación de tamaño 31, se ha obtenido un error absoluto medio de 19882,8823, un error relativo de 0,1674 y un error cuadrático medio de 681.301.217,41.

$$\begin{aligned}\widehat{Precio} = {} & 242496{,}40 - 26{,}55 \cdot Distancia + 24731{,}68 \cdot NumBaños + 555{,}48 \\ & \cdot Tamaño + 14088{,}93 \cdot Estadogood + 25697{,}23 \\ & \cdot Estadonewdevelopment - 38900{,}69 \cdot CocheraFALSE - 13023{,}83 \\ & \cdot AscensorFALSE - 13769{,}82 \cdot AireAconFALSE - 12922{,}29 \\ & \cdot TrasteroFALSE\end{aligned}$$

Además del modelo de regresión, se ha estimado una red neuronal a partir de las mismas variables. El mejor resultado se ha obtenido con una red con una única capa oculta compuesta por tres neuronas, y usando como función de activación la función logística. Las estimaciones de los pesos de la red se muestran en la Tabla 40.

Tabla 40. Estimación los pesos de la red. Viviendas. Bellavista – Jardines de Hércules.

	Valores		Valores
Error	0,5335	**Estadonewdevelopment.to.1layhid2**	0,6237
reached.threshold	0,0010	**CocheraFALSE.to.1layhid2**	-3,3515
Steps	4279	**AscensorFALSE.to.1layhid2**	0,2208
Aic	69,0671	**AireAconfalse.to.1layhid2**	-0,3740
Bic	182,3795	**TrasteroFALSE.to.1layhid2**	0,8511
Intercept.to.1layhid1	-1,7680	**Intercept.to.1layhid3**	-0,9062
Distancia.to.1layhid1	1,0978	**Distancia.to.1layhid3**	4,0097
NumBaños.to.1layhid1	-3,8276	**NumBaños.to.1layhid3**	9,9840
Tamaño.to.1layhid1	-0,7089	**Tamaño.to.1layhid3**	-16,1097
Estadogood.to.1layhid1	0,5646	**Estadogood.to.1layhid3**	3,4126
Estadonewdevelopment.to.1layhid1	-1,1075	**Estadonewdevelopment.to.1layhid3**	-354,878
CocheraFALSE.to.1layhid1	1,4489	**CocheraFALSE.to.1layhid3**	-3,6884
AscensorFALSE.to.1layhid1	0,7145	**AscensorFALSE.to.1layhid3**	-2,0010
AireAconfalse.to.1layhid1	0,2874	**AireAconfalse.to.1layhid3**	-0,4486
TrasteroFALSE.to.1layhid1	0,6109	**TrasteroFALSE.to.1layhid3**	-1,7923
Intercept.to.1layhid2	2,2502	**Intercept.to.Precio**	-0,1247
Distancia.to.1layhid2	3,3676	**1layhid.1.to.Precio**	-0,9585
NumBaños.to.1layhid2	-4,0867	**1layhid.2.to.Precio**	1,1328
Tamaño.to.1layhid2	2,8858	**1layhid.3.to.Precio**	-0,4011
Estadogood.to.1layhid2	4,1695		

Fuente: Elaboración propia

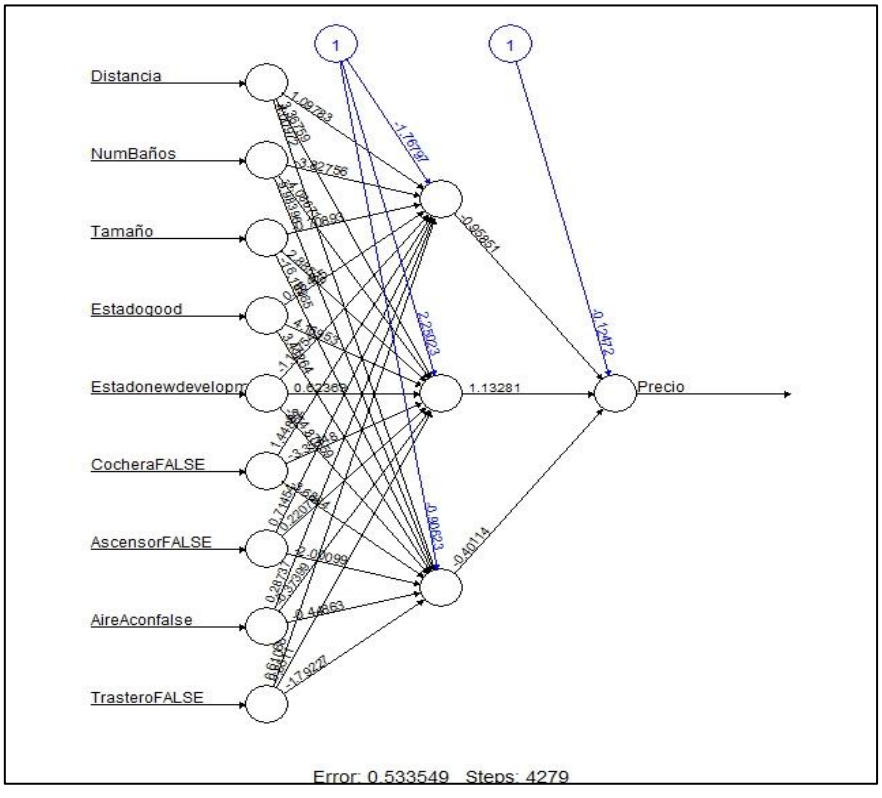


Figura 160. Arquitectura de la red. Viviendas. Bellavista – Jardines de Hércules.

Para esta red, se ha obtenido un error cuadrático medio de 609.851.640,55. La representación de la arquitectura de la red se muestra en la Figura 160.

8.2.5.2. Centro

El distrito Centro es el de mayor precio de la ciudad, hay un menor número de viviendas que disponen de piscina y el aparcamiento público es muy limitado.

El modelo se ha construido a partir de una muestra con 2.392 viviendas. El modelo final, con todas las observaciones, contiene 8 variables que son: Distancia, Tamaño, Número de baños, Estado de la vivienda y las relativas a la existencia de terraza, trastero, cochera y piscina. El valor del coeficiente de determinación es de 0,72 y el del criterio de información de Akaike 65.132,4731.

Se detectan como observaciones influyentes, aquellas viviendas que al ser eliminadas generan un valor de AIC inferior a 65.068,4648. En total se obtienen 8 inmuebles, todos con precios de 1.700.000 € o superiores.

Al eliminar estas observaciones, la muestra queda con un tamaño de 2.384, el valor del coeficiente de determinación aumenta hasta 0,7711 y el valor del criterio de información disminuye hasta 63.827,2224. Para una muestra de validación de 215 viviendas, el error relativo medio ha dado un resultado de 0,2478.

El modelo final obtenido es:

$$\widehat{Precio} = 40521{,}10 - 37{,}62 \cdot Distancia + 44601{,}70 \cdot NumBaños + 1935{,}94 \cdot Tamaño + 137861{,}15 \cdot Estadogood + 108753{,}71 \cdot Estadonewdevelopment - 31562{,}10 \cdot CocheraFALSE - 27187{,}28 \cdot TerrazaFALSE - 46531{,}43 \cdot PiscinaFALSE - 10654{,}86 \cdot TrasteroFALSE$$

A diferencia que lo que ocurría a nivel general, en el distrito Centro no es relevante, en la valoración de la vivienda, la disponibilidad de aire acondicionado, ascensor o armarios empotrados; pero sí lo es la disponibilidad de terraza y el trastero.

La poca relevancia del ascensor se justifica por el reducido número de viviendas a la venta, situadas en pisos altos, ya que casi un 80 % se sitúan en la segunda planta o una inferior. La influencia, en este caso, del trastero puede estar relacionada con un mayor precio del suelo y por tanto de cocheras y trasteros. De hecho, el incremento en el precio de la vivienda al disponer de cochera, es significativamente superior en el distrito Centro.

Nuevamente, la mejor arquitectura de red obtenida es la compuesta por una única capa oculta y con 3 neuronas en esta capa. Un menor número de neuronas dificulta la convergencia, y un mayor número de ellas devuelve peores resultados. La Figura 161, muestra los resultados de la red óptima, que mejora en casi un 19 % el error cuadrático medio obtenido por el mejor modelo de regresión.

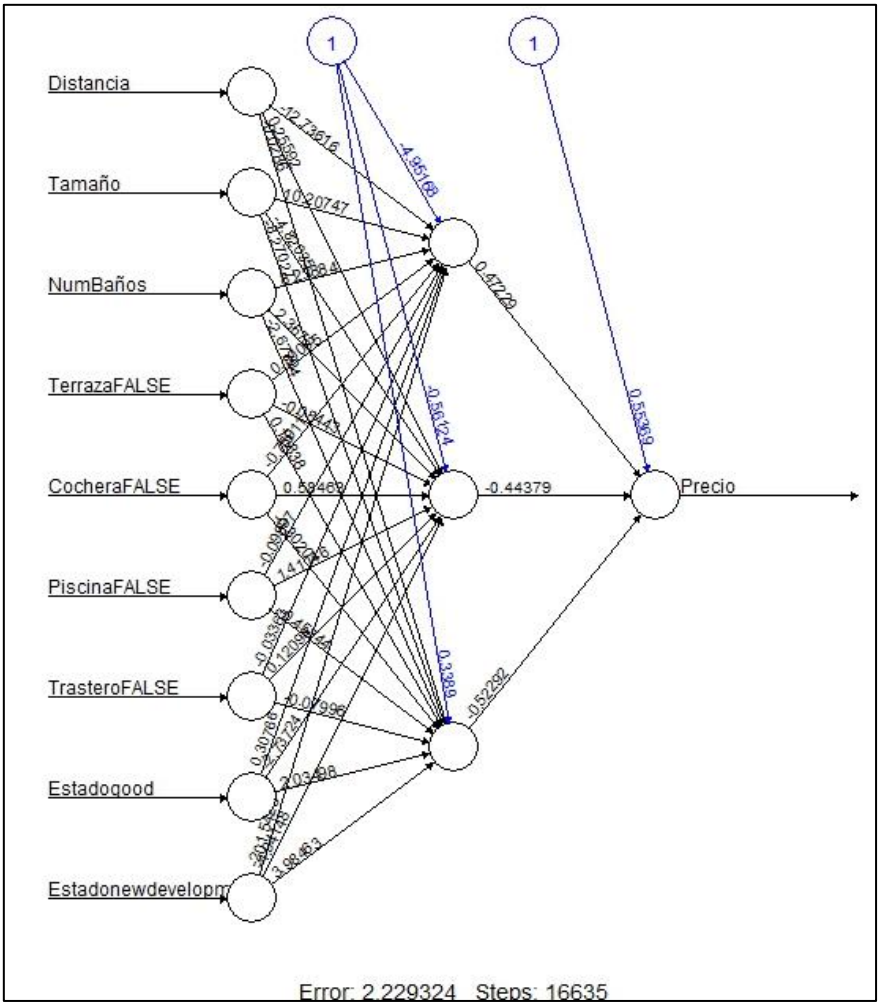


Figura 161. Arquitectura de la red. Viviendas. Centro. (Fuente: Elaboración propia)

8.2.5.3. Cerro Amate

Es, junto con Torreblanca, uno de los barrios con menor renta per cápita de la ciudad. En éste, se disponen de 1.054 viviendas a la venta.

El modelo estimado para el conjunto de datos completo incluye como variables explicativas, el tamaño, el número de baños, el estado de la vivienda y las variables que indican si la vivienda dispone de armario empotrado, piscina o ascensor. Su coeficiente de determinación es 0,6934.

Se han identificado 62 viviendas como influyentes, que son, en este caso, viviendas con precio elevado, pero también de precio excesivamente bajo en relación a su tamaño. Tras eliminar estas observaciones, para las 992 viviendas restantes, se ha construido el modelo de regresión que ha arrojado un coeficiente de determinación de 0,7637. El error relativo del modelo, para una muestra de 94 viviendas, es 0,2716, y el error absoluto medio 20.222,33.

El modelo estimado es:

$$\widehat{Precio} = 21481{,}05 + 16628{,}70 \cdot NumBaños + 688{,}85 \cdot Tamaño + 5494{,}94 \cdot Estadogood + 72634{,}09 \cdot Estadonewdevelopment - 8014{,}84 \cdot EmpotradosFALSE - 10866{,}75 \cdot PiscinaFALSE - 20432{,}98 \cdot AscensorFALSE$$

Al contrario que lo sucedido en el resto de zonas analizadas, en este distrito, la distancia al centro geográfico no influye en el precio de la vivienda. Además, el incremento de precio experimentado por cada metro cuadrado que aumenta el tamaño de la vivienda, con 688,85 €, es muy inferior al resto. El buen estado de la vivienda incrementa su precio significativamente, y la presencia de ascensor, piscina y armarios empotrados también.

La gran mayoría de las viviendas de este distrito son pisos o dúplex de bajo precio, esto provoca que el tipo de vivienda no influya en su precio. Tampoco tienen influencia significativa, características como la terraza, el aire acondicionado o el trastero.

La mejor red neuronal calculada para estas mismas variables, se ha obtenido con una capa oculta formada por dos neuronas, con un valor de AIC = 43,78 y un error cuadrático medio de 402.510.162,13, casi un 49 % menor que el obtenido para el modelo de regresión.

Su arquitectura y el valor de los pesos óptimos se muestran en la Figura 162.

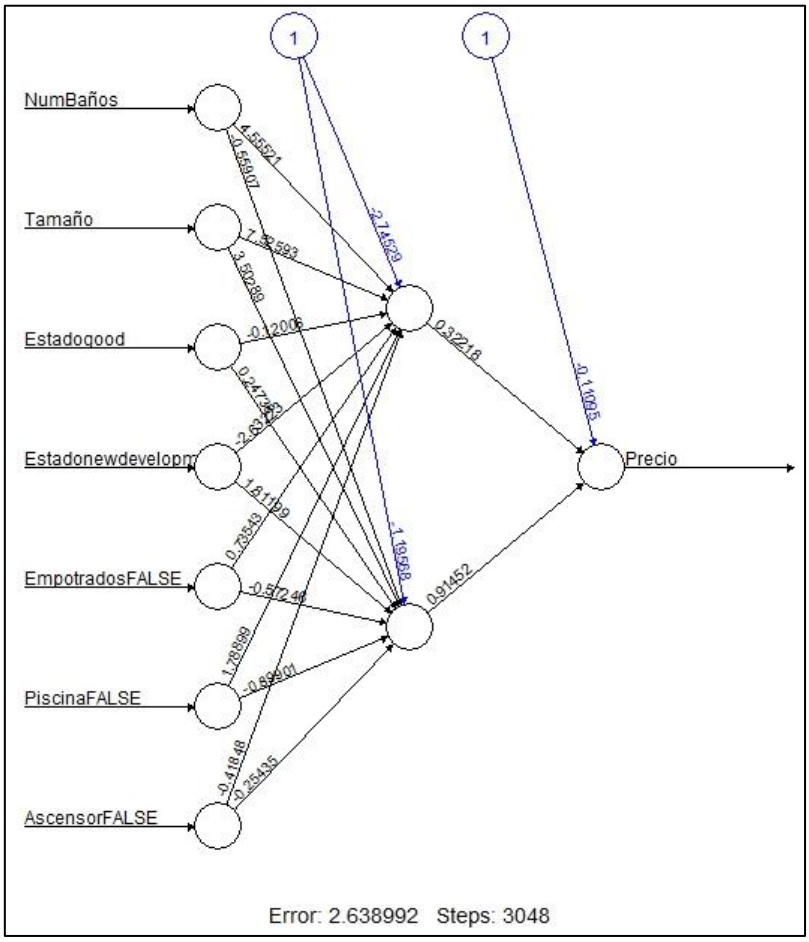


Figura 162. Arquitectura de la red. Viviendas. Cerro Amate. (Fuente: Elaboración propia)

8.2.5.4. Los Remedios

En el distrito de los Remedios, las viviendas tienen un precio medio por metro cuadrado superior a la media de la ciudad y es el segundo distrito con el precio medio de la vivienda más alto de la ciudad. Al contrario que ocurre con el barrio de Santa Clara, la muestra es suficientemente grande para hacer el análisis requerido, con 864 inmuebles.

El modelo de regresión construido con todas las observaciones tiene un coeficiente de determinación de 0,8213 y un valor del criterio de información de 22.607,0318.

Al aplicar el método de detección de observaciones influyentes, se detectan 35. Se eliminan, y para las 829 viviendas restantes, se construye el modelo un modelo de regresión con un coeficiente de determinación de 0,8441 considerando como variables explicativas el tamaño de la vivienda, la distancia al centro geográfico, el estado y la presencia de cochera, piscina y aire acondicionado. El criterio de información disminuye su valor hasta 21.994,7537. Para una muestra aleatoria de tamaño 77, el error absoluto medio calculado es 61.450,63, y el error relativo medio 0,1506.

La formulación final del modelo es:

$$\widehat{Precio} = 116677{,}91 + 5273{,}01 \cdot NumBaños + 2319{,}03 \cdot Tamaño - 26.50 \cdot Distancia + 43266{,}30 \cdot Estadogood + 84416{,}44 \cdot Estadonewdevelopment - 31827{,}99 \cdot CocheraFALSE - 101578{,}46 \cdot PiscinaFALSE - 13057{,}75 \cdot AireAconfalse$$

En esta zona vuelve a ser influyente, para determinar el precio de la vivienda, la distancia al centro geográfico, de forma que, a mayor distancia, menor precio. El sentido de la relación entre el precio y el tamaño, es contraria a la anterior, de modo que, a mayor superficie, mayor precio.

Es más, el incremento de precio unitario casi coincide con el precio medio por metro cuadrado de la vivienda de la zona, que como antes se vio, asciende a 2.291,41 €. La ausencia de cochera o aire acondicionado influye negativamente en el precio.

Es destacable el elevado peso que tiene sobre el precio de la vivienda de esta zona, la disponibilidad de piscina, que con un 95 % de confianza podemos afirmar que está comprendido entre 34.263 € y 174.576 €.

El mejor resultado obtenido de red neuronal, se ha logrado con una red con una capa oculta y una única neurona en dicha capa. El valor del criterio de información es de 21,2588 y el error cuadrático medio en una muestra de validación del mismo tamaño, se reduce un 17,89 %.

A continuación, se muestra la arquitectura de esta red junto a los pesos óptimos de ésta (Véase *Figura 163*).

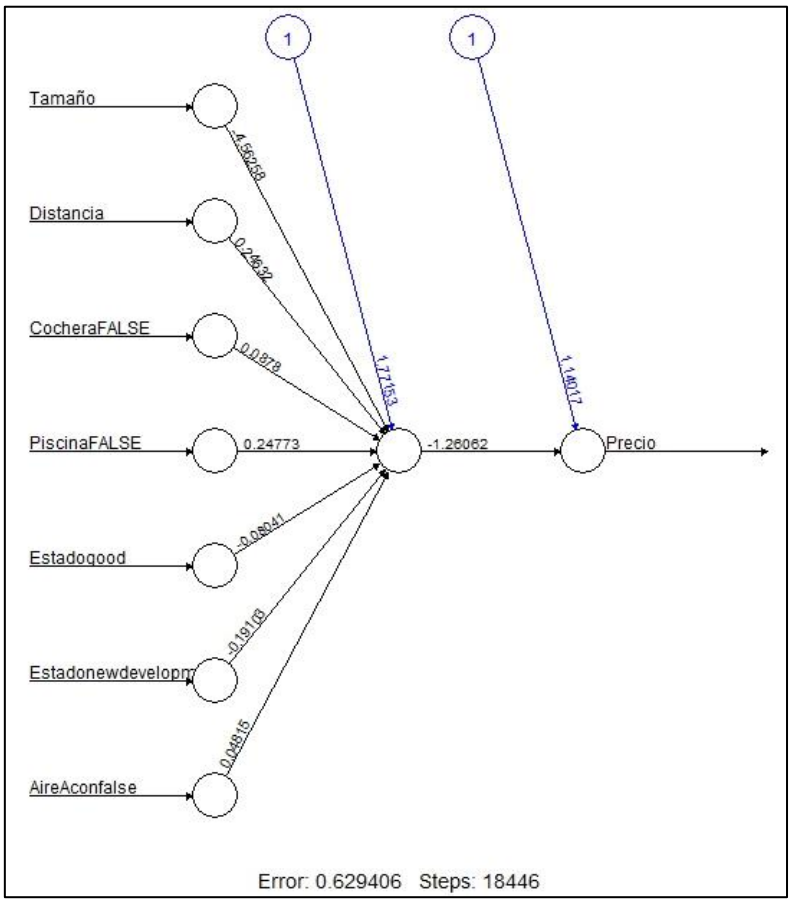


Figura 163. Arquitectura de la red. Viviendas. Los Remedios. (Fuente: Elaboración propia)

8.2.5.5. Macarena

El barrio de Macarena se encuentra al norte del distrito centro, y tiene un precio medio por debajo de la media. Es un barrio obrero con numerosos edificios de viviendas, que se extiende por una zona amplia de la ciudad.

En éste se han considerado 710 viviendas, para las que se han construido los modelos de estimación del precio de oferta.

El modelo de regresión original, se ha construido con las variables explicativas: distancia al centro geográfico, tamaño, número de fotografías del anuncio, número de baños y las referentes a disponibilidad de cochera, ascensor, piscina, climatización,

terraza y armarios empotrados. El coeficiente de determinación obtenido es de 0,8055 y el valor del criterio de información 17.144,8431.

Posteriormente se han identificado 31 observaciones influyentes que han sido eliminadas de la muestra, dejando ésta con un total de 679 viviendas.

Una vez filtrada la muestra deseada el modelo de regresión final obtenido arroja un coeficiente de determinación que indica que éste explica un 84,11 % de la variabilidad en el precio del inmueble. El valor del criterio de información de Akaike es 16.208,3665.

La formulación del modelo estimado se muestra a continuación:

$$\widehat{Precio} = 228439{,}83 + 36332{,}79 \cdot NumBaños + 1028{,}15 \cdot Tamaño - 31{,}05 \cdot Distancia + 553{,}1805 \cdot NumFotos - 18944{,}49 \cdot CocheraFALSE - 9938{,}81 \cdot AscensorFALSE - 133743{,}06 \cdot PiscinaFALSE - 8097{,}23 \cdot AireAconfalse - 6285{,}96 \cdot TerrazaFALSE - 14108{,}76 \cdot EmpotradosFALSE$$

En este modelo se ha encontrado, por primera vez, que una variable que no es una característica de la vivienda, como es el número de fotografías del anuncio en el que ésta se oferta, influye en el precio de ésta, de forma que los anuncios de viviendas de mayor precio disponen de una cantidad más alta de fotografías. Con un 95 % de confianza, el valor de la vivienda es mayor entre 210,49 € y 895,87 € por cada fotografía adicional que tenga el anuncio.

De las variables que indican la existencia de una característica adicional en la vivienda, vemos que la que más influencia tiene sobre el precio es la existencia de piscina, de forma mucha más significativa que el resto.

El incremento de un metro cuadrado supone un aumento en el precio de entre 910,16 € y 1146,15 €, algo por encima de los 1241 € de precio medio de la zona.

Para una muestra aleatoria de validación de 61 inmuebles, el error absoluto medio es 20.790,35 y el error relativo medio 0,246.

De nuevo, en esta ocasión, la mejor red que se ha podido construir tiene una capa oculta y una única neurona en esta capa (Véase *Figura 164*). Con ella se ha conseguido

reducir un 30,7 % el error cuadrático medio obtenido para el modelo de regresión. En la se muestra su arquitectura.

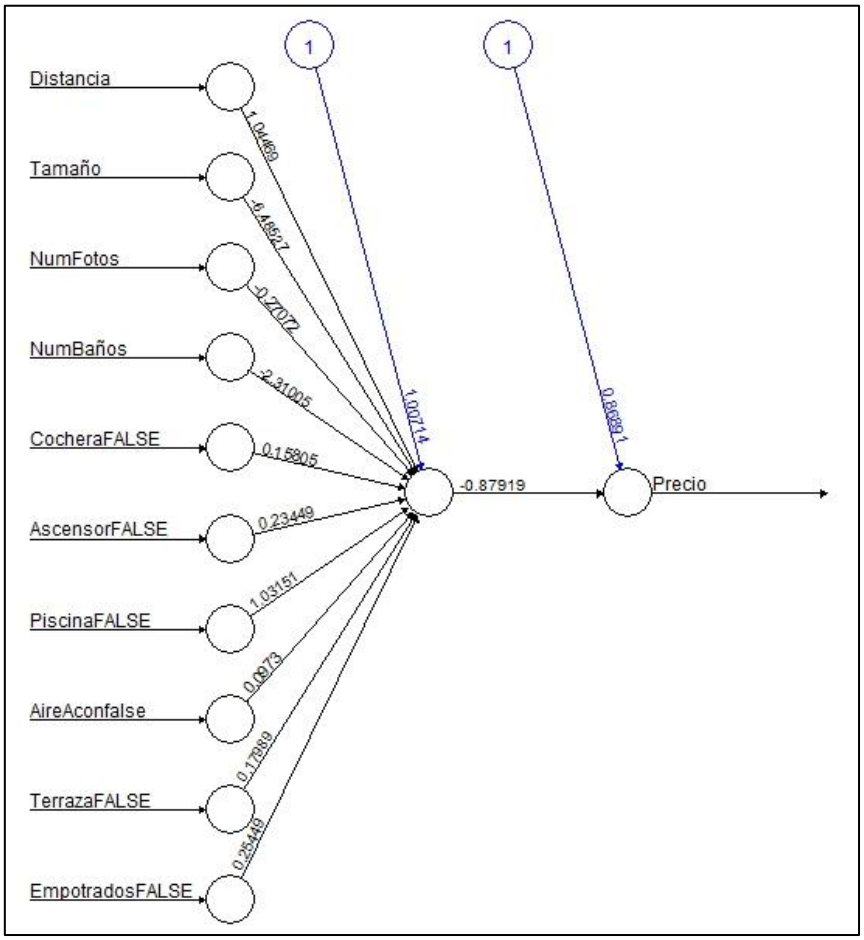


Figura 164. Arquitectura de la red. Viviendas. Macarena. (Fuente: Elaboración propia)

Como hemos visto, las principales variables que influyen en la determinación del precio de la vivienda son, por un lado, el tamaño y su situación geográfica, y por otro las características adicionales de las que dispone como su estado, el número de baños, que está fuertemente relacionado con el número de habitaciones, y la disponibilidad de ascensor, cochera o piscina.

Estas características difieren en las distintas zonas de la ciudad según la tipología de la zona en la que se encuentra. Por ejemplo, la cochera no es tan valorada en barrios con grandes espacios públicos y gratuitos de aparcamiento, pero si lo es en zonas céntricas.

8.3. Mercado de inmuebles comerciales de la aglomeración de Sevilla

El mercado inmobiliario es un sector ampliamente analizado por el peso que tiene en la economía de un país. La mayor parte de las transacciones inmobiliarias que se realizan, se corresponden con inmuebles destinados a la vivienda. No obstante, el mercado de inmuebles vinculados al desarrollo de una actividad comercial e industrial nos ofrece una visión de la configuración del sector terciario en una zona concreta.

El área metropolitana de Sevilla es uno de los núcleos poblacionales, económicos e industriales más importantes de España. Reúne a más de un millón y medio de personas, y está compuesto por la ciudad de Sevilla y una multitud de pequeños municipios que se encuentran a escasa distancia del centro de la ciudad.

Es importante poner de manifiesto las claras diferencias estructurales entre los locales comerciales y las naves industriales, debido a sus distintos usos. Mientras que la situación geográfica de un local comercial es indispensable para un buen funcionamiento de la actividad, las naves industriales suelen situarse en la periferia de las ciudades.

Asimismo, los locales comerciales son inmuebles que generalmente se sitúan en la planta baja de un edificio de viviendas y que su principal uso es el comercio. Existe una gran variedad de tamaños y forma de la planta del local, pero en general suelen ser de tamaño pequeño, la mayor parte, inferiores a 150 metros cuadrados. La visibilidad del local es fundamental, por lo que cobra especial importancia que esté situado a pie de calle o en el interior de un centro comercial, que tenga una gran fachada y, de ser posible, que esté situado en la esquina entre dos calles, para aumentar aún más, su visibilidad.

Por su parte, las naves industriales suelen estar ubicadas en zonas de fácil acceso de vehículos de gran tamaño, en lugares con buena conexión vial y en los que el uso es exclusivamente industrial. El tamaño del inmueble tiene especial importancia ya que, en general, se requieren grandes zonas de almacenaje y la maquinaria a utilizar es de grandes dimensiones. Suelen ocupar el espacio de toda la construcción y tener un reducido lugar reservado para el trabajo de oficina. Su elevado tamaño implica que el precio de venta de estos inmuebles suele ser mayor.

La aplicación de la metodología de precios hedónicos en el mercado de los locales comerciales es muy escasa y concretamente para naves industriales no se ha encontrado precedente alguno acerca de la citada metodología. El hecho de asociar, en numerosas publicaciones, mercado inmobiliario en general, con el mercado de la vivienda, da una idea del superlativo peso que éste tiene. Esto hace olvidar la importancia que un completo conocimiento del submercado de locales comerciales y naves industriales tiene también dentro de este sector.

Se ha analizado la oferta de locales comerciales y naves industriales, a la venta a 30 de octubre de 2016, en la ciudad de Sevilla y los municipios colindantes que se encuentran, a lo sumo, a una distancia de 15 kilómetros del centro. Por tanto, además de los propios de la capital sevillana, se han incluido inmuebles de los municipios: Albaida del Aljarafe, Alcalá de Guadaira, Alcalá del Río, Almensilla, Bollullos de la Mitación, Bormujos, Camas, Castilleja de Guzmán, Castilleja de la Cuesta, Coria del Río, Dos Hermanas, Espartinas, Gelves, Gines, La Algaba, La Puebla del Río, La Rinconada, Mairena del Aljarafe, Montequinto – en Dos Hermanas -, Olivares, Palomares del Río, Salteras, San José de la Rinconada - La Rinconada -, San Juan de Aznalfarache, Sanlúcar la Mayor, Santiponce, Tomares, Umbrete, Valencina de la Concepción y Villanueva del Ariscal.

La muestra está formada por un total de 1916 inmuebles, que es el total de inmuebles a la venta en el portal idealista.com, en la fecha indicada. Este portal es el portal web inmobiliario más importante a nivel nacional, debido a la extensa oferta de que dispone, tanto a la venta como en alquiler.

La oferta de inmuebles publicados en el portal idealista.com, proviene de inmobiliarias, bancos e incluso particulares, por lo que esta es la procedencia de los locales comerciales analizados. La muestra está compuesta por un total de 1236 locales comerciales y naves industriales en la ciudad de Sevilla, y 680 en los municipios colindantes -de los que el 44% se concentra en Dos Hermanas, Mairena del Aljarafe y Alcalá de Guadaira-.

De todos ellos, 1506 – el 78,6% - son locales comerciales, de los que 1056 están situados en la ciudad de Sevilla y el resto, 450, están situados en los municipios del extrarradio. Respecto a las naves industriales, con 410, representan el 21,4% de la oferta

total de inmuebles analizados, de los que 180 están situados en la ciudad de Sevilla, y 230 en los municipios del extrarradio -un 56,09% del total de naves industriales-. Esto es indicativo de la presencia de una gran actividad industrial en los municipios cercanos a la ciudad de Sevilla debido a un menor precio del suelo en el extrarradio, que unido al elevado espacio que se necesita para llevar a cabo esta actividad, y a la cercanía al principal núcleo de población, hacen de éste, un enclave privilegiado.

Las características analizadas son:

- Precio de venta del inmueble.
- Tamaño en metros cuadrados.
- Número de baños.
- Tipo de inmueble: local comercial o nave industrial.
- Número de fotografías[3] del anuncio en el portal.
- Barrio.
- Municipio.
- Distrito.
- Barrio.
- Piso en el que se encuentra el inmueble.
- Si dispone de aire acondicionado.
- Localización. Los valores posibles son: A pie de calle, entreplanta, en el interior de un centro comercial u otro.
- Si se encuentra haciendo esquina.
- Si tiene salida de humos.

Respecto a la variable "distrito" se ha considerado, al igual que en el caso anterior, la división efectuada en el portal fuente de los datos, por lo que los distritos recogidos, que se muestran en la Figura 165, han sido: Bellavista – Jardines de Hércules, Centro, Cerro Amate, La Palmera – Los Bermejales, Los Remedios, Macarena, Nervión, Parque Alcosa, Pino Montano, Prado de San Sebastián – Felipe II, San Jerónimo, San Pablo, Santa Clara, Santa Justa – Miraflores – Cruz Roja, Sevilla Este, Torreblanca y Triana. Algunos de los

[3] Aunque no es una característica propia del inmueble, ésta es determinante para su divulgación. Lo qiue la hace determinante en la decisión de compra y en el período de tiempo transcurrido hasta la misma.

11 distritos de la ciudad de Sevilla han sido divididos por su extensión como es el caso del distrito Norte.

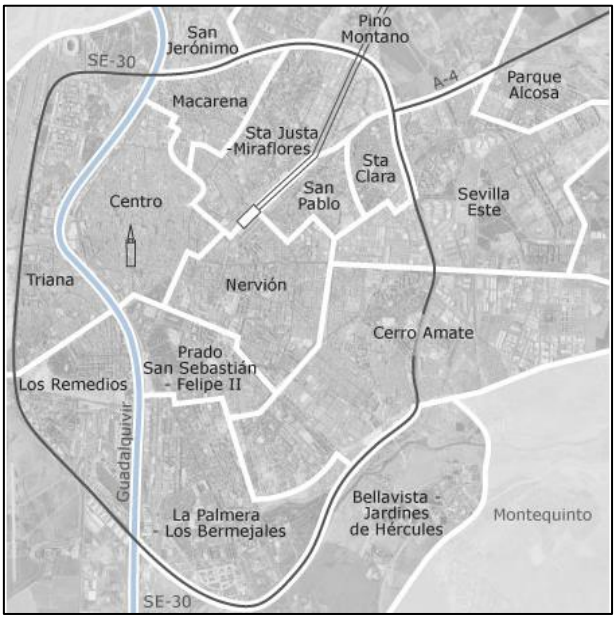


Figura 165. Distritos de la ciudad de Sevilla. (Fuente: idealista.com)

En una primera fase del estudio, se ha realizado un análisis descriptivo de las características de los inmuebles analizados, comparando los que se encuentran en la ciudad de Sevilla con los que se encuentran en los municipios de su alrededor, distinguiendo a su vez entre los locales comerciales y las naves industriales, por sus características diferenciadoras.

A continuación, se ha construido un modelo para la estimación del precio de venta de las naves industriales de la ciudad de Sevilla, por un lado, y de los municipios de su entorno, por otro, utilizando la metodología de precios hedónicos, por lo que se estimará el precio a partir de las características del inmueble.

8.3.1. Análisis descriptivo

En la ciudad de Sevilla, se han encontrado locales comerciales para los que su precio oscila entre los 6.900 € de un local comercial en Torreblanca de 12 metros cuadrados y los 6.500.000 € de un edificio comercial de 4 plantas situado en una de las calles más comerciales de la ciudad. No obstante, el 95% de los inmuebles analizados tienen un precio inferior a los 750.000 €, y el precio medio se sitúa en 267.331,26 €.

Por tamaño, el local más pequeño está situado en el distrito de Pino Montano, y tiene una extensión de 11 metros cuadrados. El de mayor tamaño, 3030 metros cuadrados está situado en Sevilla Este y es un antiguo concesionario de coches. El tamaño medio de los locales comerciales de la ciudad de Sevilla se sitúa en 168 metros cuadrados.

Si observamos el precio por metro cuadrado, el más barato es de tan sólo 97,15 € por metro cuadrado correspondiente a un local subterráneo de 772 metros cuadrados, que realmente es un garaje, situado en Cerro Amate. El más caro, con un precio de 11.818,18 € por metro cuadrado, se corresponde con el local comercial más caro de la ciudad, antes comentado. 1657,67 € es el precio medio por metro cuadrado para la ciudad de Sevilla.

Al analizar las naves industriales de la ciudad, como cabe esperar, el tamaño y el precio medio, con 873,32 metros cuadrados y 473.255,60 € respectivamente, son significativamente superiores a los de los locales comerciales. No obstante, el precio medio por metro cuadrado es claramente inferior al de los locales comerciales, con un valor de 637,52 €.

En los municipios colindantes, el local de menor precio encontrado, está situado en el casco histórico de Mairena del Aljarafe de 65 metros cuadrados, con un precio de venta de 6.000 €. El de mayor precio se encuentra en Tomares, y es un local comercial de más de 1.000 metros cuadrados y un precio de venta de 2.242.664 €. El 95% de los locales comerciales y las naves de esta zona tienen un precio por debajo de los 430.000 € y el precio medio es de 146.247,84 € -muy por debajo del precio medio de la ciudad de Sevilla-.

El tamaño de los locales de estos municipios oscila entre los 20 de un local de alimentación en San José de la Rinconada, y los 2.683 de un local en Olivares. No obstante, el tamaño medio es de 174,29 metros cuadrados, ligeramente superior al calculado para la capital de la provincia.

El precio medio por metro cuadrado de los locales de los municipios próximos es un 40% inferior al precio de Sevilla, y se sitúa en 986,98 €. El precio más bajo se corresponde con el local de Mairena del Aljarafe antes comentado, y el más elevado pertenece a un pequeño local de la zona comercial de San Juan de Aznalfarache, con un precio de 3.214,29 € por metro cuadrado.

Por distritos de la ciudad de Sevilla, Nervión y Centro son los que aportan un mayor número de locales a la muestra, y Cerro Amate el que más naves industriales contiene. A través de un análisis de la varianza, no se han hallado diferencias significativas en el tamaño medio de los locales comerciales de los distintos distritos, pero sí en el precio total y el precio por metro cuadrado. En el primero de ellos, a través del test de Duncan, observamos, que en el distrito Centro, el precio medio es superior al del resto. Respecto al precio por metro cuadrado, los distritos en los que éste es más económico, son Torreblanca y Cerro Amate; y los de precio más elevado son, nuevamente Centro, Los Remedios y Santa Clara.

En las naves industriales también se hallan diferencias significativas en el precio y el precio unitario medio. Hay grandes diferencias en los precios totales medios, mientras que en San Jerónimo es de 118.802,38 €, en Sevilla Este el valor se sitúa en 1.133.692,31 €. Sin embargo, es en Torreblanca donde el precio por metro cuadrado es significativamente inferior con 520,35 €, mientras que en Prado de San Sebastián – Felipe II es casi el doble, 978,87 €.

En la Tabla 41 se muestra el precio medio total y por metro cuadrado, el tamaño medio, así como el intervalo de confianza para el precio medio de los locales comerciales de los distintos distritos de la ciudad de Sevilla. Se han omitido los resultados correspondientes al distrito de Santa Clara por la escasa oferta.

Se han omitido los resultados correspondientes al distrito de Santa Clara por la escasa oferta.

Respecto a las características de los inmuebles comerciales de la ciudad de Sevilla, podemos destacar que el 36,7% de las naves industriales y el 27% de los locales, no dispone de aseos y un 32,2% de las naves y casi un 46% de los locales dispone de tan sólo uno, por lo que, en contra de lo que cabía esperar, el porcentaje de locales con un número deficiente de aseos es superior que el de naves industriales.

Sólo el 36% de los locales comerciales y el 25% de las naves industriales a la venta de la ciudad de Sevilla disponen de aire acondicionado, en un lugar donde la climatología hace de éste un elemento indispensable.

Tabla 41. Precio y tamaño de los locales comerciales. Ciudad de Sevilla.

	Tamaño muestral	Precio Medio	Ext. inf. IC Precio Medio 90 %	Ext. sup. IC Precio Medio 90 %	Tamaño Medio	Precio medio por metro cuadrado
Bellavista – Jardines de Hércules	36	166.531,39 €	110.488,80 €	222.573,97 €	118,83	1.242,39 €
Centro	169	518.329,31 €	397.137,93 €	639.520,68 €	205,69	2.485,28 €
Cerro Amate	105	159.568,82 €	131.090,71 €	188.046,93 €	172,74	1.069,91 €
La Palmera - Los Bermejales	28	236.360,71 €	169.600,23 €	303.121,20 €	141,64	1.780,35 €
Los Remedios	62	391.832,94 €	263.426,26 €	520.239,61 €	235,16	1.911,11 €
Macarena	69	185.925,72 €	149.650,05 €	222.201,40 €	120,33	1.561,68 €
Nervión	181	249.747,85 €	215.562,56 €	283.933,13 €	163,75	1.613,22 €
Parque Alcosa	11	128.537,64 €	82.717,27 €	174.358,00 €	106,27	1.228,33 €
Pino Montano	22	117.121,91 €	86.160,25 €	148.083,56 €	103,45	1.349,50 €
Prado de San Sebastián - Felipe II	33	259.553,33 €	162.724,62 €	356.382,05 €	182,70	1.589,31 €
San Jerónimo	32	155.223,88 €	55.203,00 €	255.244,75 €	159,41	1.237,87 €
San Pablo	12	118.829,58 €	67.335,23 €	170.323,94 €	79,58	1.713,75 €
Santa Justa - Miraflores - Cruz Roja	83	189.252,82 €	144.488,34 €	234.017,29 €	129,24	1.620,61 €
Sevilla este	94	229.555,47 €	122.833,37 €	336.277,57 €	179,90	1.489,83 €
Torreblanca	11	143.788,09 €	68.460,90 €	219.115,28 €	204,82	1.046,39 €
Triana	106	231.723,44 €	195.473,04 €	267.973,85 €	120,43	1505,87 €

Fuente: Elaboración propia

Menos de un 1% de la oferta de locales comerciales se encuentra en el interior de un centro comercial y sólo el 13% están situados haciendo esquina entre dos calles. Un porcentaje algo mayor, el 16%, dispone de salida de humos.

El análisis de los locales comerciales del extrarradio, en función del municipio en el que estos se encuentran, - Tabla 41- arroja que San Juan de Aznalfarache es el municipio en el que el precio medio es menor -con algo menos de 100.000 € de media-, mientras que en Tomares se encuentra el precio medio más elevado que asciende a 248.380,63 €. El bajo precio de los locales de San Juan de Aznalfarache se explica por el reducido tamaño de los locales de este municipio, mientras que el elevado precio medio en Tomares se justifica por ser uno de los términos en los que el tamaño medio de los inmuebles es el más elevado y también por un elevado precio del suelo en el municipio con mayor renta per cápita de Andalucía.

No obstante, el precio por metro cuadrado más elevado y el tamaño medio más bajo se encuentran en Montequinto, con un valor medio de 1.762,99 € y 67,80 metros cuadrados respectivamente. El precio medio por metro cuadrado más bajo -con algo menos de 650 €- se sitúa en Gelves.

En las naves industriales se aprecian diferencias significativas en el tamaño y el precio por metro cuadrado en los distintos municipios, de forma que el tamaño medio más bajo se encuentra en Gines -con 347,64 metros cuadrados-, cinco veces más pequeño que el encontrado en Dos Hermanas -1.834,63 metros cuadrados-. El precio por metro cuadrado de las naves de Mairena del Aljarafe - con casi 700 € - es muy superior al resto de los municipios analizados -que se sitúan entre los 411 € y los 500 €-.

Tabla 42. Precio y tamaño de locales comerciales. Municipios próximos.

	Tamaño muestral	Precio Medio	Ext. inf. IC Precio Medio 90 %	Ext. sup. IC Precio Medio 90 %	Tamaño medio	Precio medio por metro cuadrado
Bollullos de la Mitación	14	146.386,29 €	7.072,35 €	285.700,22 €	205,36	793,96 €
Bormujos	28	128.040,18 €	82.988,91 €	173.091,44 €	161,46	828,34 €
Camas	24	160.686,13 €	76.149,80 €	245.222,45 €	218,33	1.010,58 €
Castilleja de la Cuesta	22	177.465,27 €	128.788,58 €	239.064,42 €	199,09	1.022,51 €
Coria del Río	14	141.999,43 €	95.967,69 €	188.031,17 €	175,07	901,09 €
Dos Hermanas	94	150.344,81 €	114.588,97 €	186.100,65 €	171,62	995,56 €
Gelves	15	106.537,27 €	72.007,26 €	141.067,28 €	169,87	648,54 €
Gines	23	105.134,83 €	68.132,60 €	142.137,05 €	155,74	903,03 €
Mairena del Aljarafe	84	140.158,36 €	104.901,53 €	175.415,19 €	165,80	919,55 €
Montequinto	15	120.352,13 €	58.220,00 €	182.483,96 €	67,80	1.762,99 €
San Jose de la Rinconada	21	131.691,24 €	55.345,12 €	208.037,36 €	162,05	845,91 €
San Juan de Aznalfarache	24	97.273,88 €	49.326,85 €	145.220,90 €	92,88	1.365,80 €
Tomares	27	248.380,63 €	79.880,92 €	416.880,34 €	193,11	1.215,03 €

Fuente: Elaboración propia

La Tabla 42 recoge la información relativa a precios y tamaños medios de los municipios cercanos a la ciudad de Sevilla en los que la oferta es suficientemente grande para unos resultados significativos.

El 77,8% de las naves industriales y el 81,5% de los locales comerciales de los municipios del extrarradio dispone de un baño o menos. El dato relativo al número de inmuebles con aire acondicionado es incluso peor en el extrarradio, dado que sólo el 22% de las naves y el 26,5% de los locales disponen de éste.

Sin embargo, el porcentaje de locales situados en el interior de un centro comercial se eleva hasta el 4%, debido a la dispersión de éstos en las zonas periféricas de la ciudad.

En términos generales, se puede afirmar, también, que existen diferencias significativas en el precio medio por metro cuadrado de un local comercial según si el inmueble posee o no aire acondicionado, teniendo un mayor precio si dispone de éste. También es significativo el aumento en el precio de los locales situados en la calle respecto a los que están situados en centros comerciales o en entreplanta, de los que tienen salida de humos respecto a los que no, y de los que están situados en la intersección de dos calles frente a los que no. Estas características no son importantes, sin embargo, al proponer el precio de las naves industriales.

Al observar el anuncio del inmueble en el portal y considerar los recursos utilizados para fomentar la venta, se observa que, aunque hay anuncios que tienen hasta 40 fotografías, hay un 9% que no dispone de fotografías y más de la mitad disponen de 5 fotografías o menos. Además, un número muy reducido de ofertas, disponen de video. Es interesante destacar que el número de fotografías de los anuncios de locales comerciales de la ciudad de Sevilla es significativamente superior al de los locales del extrarradio.

En la *Figura 166* se muestra la posición de los locales comerciales analizados, sobre una imagen satélite, así como la información relativa al precio por metro cuadrado de éstos a través de la escala de colores que se puede consultar en la parte derecha de la imagen. Cada punto representado en el mapa muestra la ubicación de un local comercial, de forma que, con un degradado de color, de amarillo a rojo, se indican los inmuebles de menor a mayor precio por metro cuadrado. Se han eliminado los inmuebles con precio por metro cuadrado superior a 4.000 € por representar datos extremos.

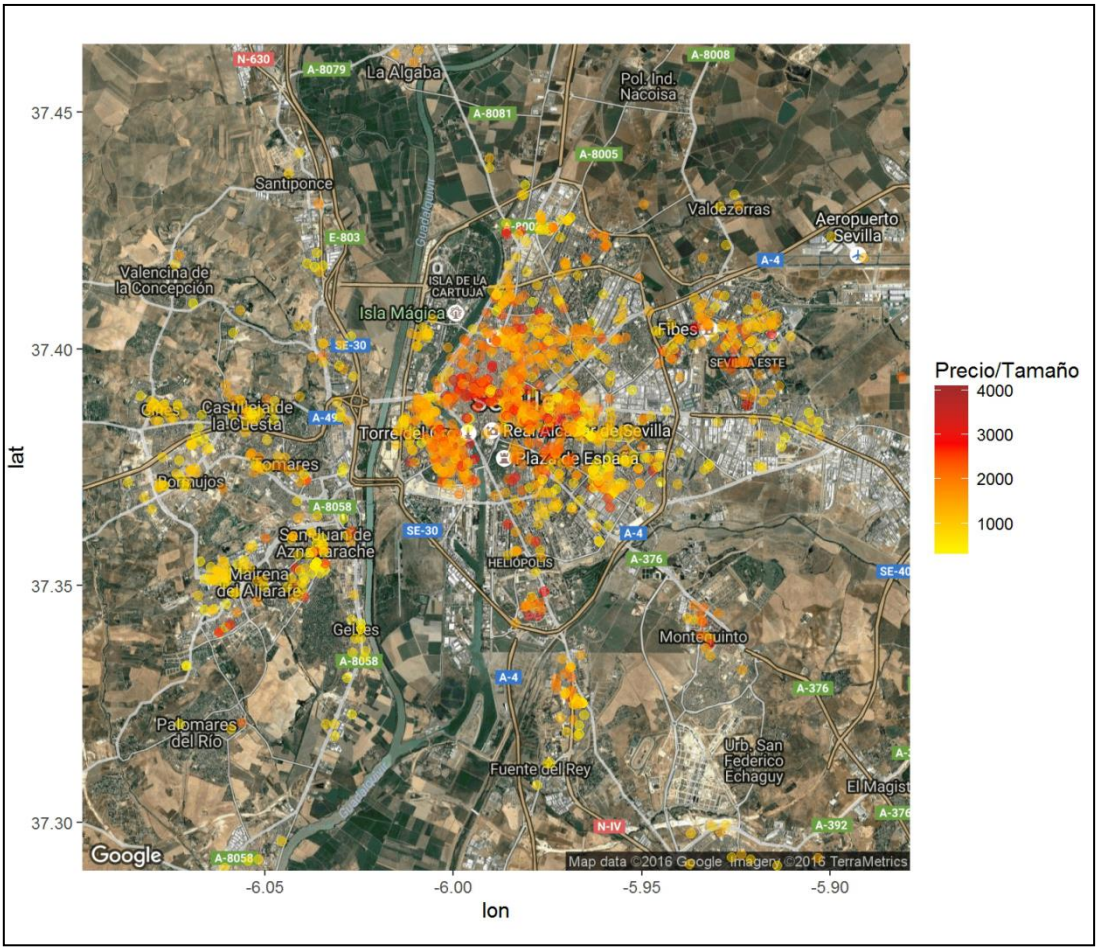


Figura 166. Mapa Sevilla. Precio por metro cuadrado de locales (Fuente: Elaboración propia)

Como puede observarse, los locales comerciales con mayor precio están situados principalmente en el municipio de Sevilla, y más concretamente en los distritos centro, La Palmera – Bellavista, Triana, Nervión y Sevilla Este. En los municipios de alrededor, podemos destacar Montequinto y, en menor medida, Tomares.

8.3.2. Estimación del precio

A continuación, se han estimado dos modelos de regresión para la estimación del precio a través de la metodología de precios hedónicos. El primero de ellos, para las naves industriales situadas en los municipios colindantes de la capital de Sevilla, y un segundo modelo para el mismo tipo de inmuebles ubicados en la propia capital.

Debido a la escasa presencia de trabajos en los que se utiliza la metodología de precios hedónicos para la estimación de precios de locales comerciales y naves industriales, se

optó por la realización de un análisis previo de la influencia que cada una de las variables consideradas tiene sobre el precio.

En primer lugar, se analizó la existencia de relación lineal entre cada una de las variables explicativas – tamaño, distancia al centro del municipio y número de baños – y el precio del inmueble. Para ello se realizaron los correspondientes contrastes sobre la nulidad del coeficiente de correlación lineal de Pearson.

La influencia que las distintas modalidades de las variables cualitativas tienen sobre el precio fue objeto del siguiente análisis. Para ello, se realizaron pruebas t de comparación de medias para las de tipo dicotómico, y contrastes de análisis de la varianza para aquellas con más de dos categorías. Primero, se estudió si el tipo de inmueble – local o nave – influye -como parece evidente por sus características diferenciadoras- en el precio. Una vez confirmada esta relación, las variables estudiadas para ambos tipos de inmuebles fueron planta, esquina, localización, salida de humos y aire acondicionado.

Por último, se justificó la necesidad de incluir las características de la zona en la que el inmueble está situada, en la construcción del modelo de precios hedónicos. En el caso de los locales comerciales de la ciudad de Sevilla se optó por comprobar la existencia de diferencias significativas entre los distintos distritos y en el caso de las naves industriales no situadas en la ciudad de Sevilla, el municipio en el que está situada.

Los resultados de este análisis previo son los siguientes:

Un análisis preliminar de la existencia de relación lineal entre el precio y las variables exógenas de carácter cuantitativo – distancia al centro, tamaño y número de baños – reveló los resultados que se muestran en la Tabla 43.

Como puede observarse, en el caso de los locales comerciales de la ciudad de Sevilla, todas las relaciones son significativas, salvo la existente entre el número de baños y la distancia al centro del municipio, cuyo resultado no es concluyente. Cuando estas relaciones las medimos con la transformada logarítmica del precio la significatividad de las relaciones se mantiene. La posible existencia de multicolinealidad entre tamaño, número de baños y distancia al centro debe considerarse en la validación del modelo, aunque los coeficientes de correlación son reducidos.

Tabla 43. Relaciones entre variables cuantitativas.

		Precio vs Tamaño	Precio vs n.º de baños	Precio vs distancia	Tamaño vs nº de baños	Tamaño vs distancia	N.º de baños vs distancia
Locales comerciales Sevilla	r	**0,703**	**0,088**	**-0,182**	**0,107**	**-0,078**	-0,057
	Prob. Límite	< 0,0001	0,0042	<0,0001	0,0005	0,011	0,066
Naves industriales extrarradio	r	**0,885**	0,075	0,134	0,046	**0,233**	0,016
	Prob. Límite	< 0,0001	0,259	0,043	0,491	< 0,0001	0,814

Fuente: Elaboración propia

Sin embargo, en el caso de las naves industriales del extrarradio, podemos afirmar que sólo el tamaño tiene una relación significativa con el precio. El resultado respecto a la distancia no es concluyente, aunque podemos afirmar la existencia de relación lineal entre ambas variables, con un nivel de significación del 5%.

Por otro lado, se ha analizado la influencia que las diferentes modalidades de las variables cualitativas tienen sobre el precio del inmueble, mediante contrastes de comparación del valor medio del precio mediante pruebas t y ANOVA según el número de modalidades.

Como se ha comentado anteriormente, el precio medio del inmueble difiere entre locales y naves industriales. Un contraste de comparación de medias confirma que estas diferencias son significativas con una probabilidad límite inferior a 0,0001.

Los resultados obtenidos para el resto de variables cualitativas se muestran en la Tabla 44.

El precio medio de los locales comerciales difiere significativamente en función de la planta en la que se sitúe. La presencia de aire acondicionado se revela también significativo para niveles de significación mayores o iguales al 5%. Aunque la localización no es relevante, al realizar este estudio con la transformación logarítmica del precio, sí ha resultado influyente. En el caso de las naves industriales, ninguna de estas variables influye significativamente en el precio.

Tabla 44. Influencia de las distintas características en el precio.

Influencia sobre el precio		**Aire acond.**	**Salida humos**	**Localización[4]**	**Esquina**	**Planta**
Locales comerciales Sevilla	t / F	1,974	-0,672	-0,162	1,121	8,879
	Prob. Límite	**0,049**	0,502	0,871	0,263	< **0,0001**
Naves industriales extrarradio	t / F	0,203	0,048	-	-0,815	0,547
	Prob. Límite	0,840	0,962	-	0,418	0,651

Fuente: Elaboración propia

Por último, se ha realizado análisis por distritos de los inmuebles de la ciudad de Sevilla, que revela que Nervión y Centro son los que aportan un mayor número de locales a la muestra, y Cerro Amate el que más naves industriales contiene. Un contraste de análisis de la varianza, ha confirmado que, si bien, no existen diferencias significativas en el tamaño medio de los locales comerciales de los distintos distritos, sí se observan en el precio total y el precio por metro cuadrado. En el primero de ellos, a través del test de Duncan, observamos, que en el distrito Centro, el precio medio es superior al del resto. Respecto al precio por metro cuadrado, los distritos en los que éste es más económico, son Torreblanca y Cerro Amate; y los de precio más elevado son, nuevamente Centro, Los Remedios y Santa Clara.

También se han observado diferencias significativas al comparar el precio por metro cuadrado de las naves industriales en función de los distintos municipios próximos a Sevilla y contenidos en la aglomeración urbana, aunque no en el precio total ni en el tamaño de éstas. Se aprecian diferencias significativas en el tamaño y el precio por metro cuadrado en los distintos municipios.

[4] No se ha podido analizar la asociación entre localización y precio debido a que todas las naves industriales están situadas a pie de calle.

De todo ello, se deduce la necesidad de incluir el distrito al que pertenece un inmueble en el modelo de estimación de su precio. Los resultados de ambos estudios se muestran en la Tabla 45.

Tabla 45. Comparación de precio, tamaño y precio / m^2 en función del municipio y el distrito

	Locales comerciales vs. distritos		Naves industriales vs. municipios	
	F	Sig.	F	Sig.
Tamaño	1,452	,110	1,459	,088
Precio	5,943	,000	,943	,541
Precio_por_m2	12,342	,000	6,747	,000

Fuente: Elaboración propia

Tras este análisis, puede concluirse que, en el caso de los locales comerciales, las variables explicativas que pueden explicar el precio del inmueble son: tamaño, número de baños, distancia, aire acondicionado, planta y distrito.

Por otro lado, las variables a considerar en la estimación del precio de las naves industriales son tamaño, distancia y municipio, que, aunque no es relevante en el precio, sí lo es en el precio unitario. Se han considerado los municipios de Alcalá de Guadaira, Bollullos de la Mitación, Camas, Dos Hermanas, Mairena y otros – en los que se han agrupado los restantes inmuebles - debido a la escasa disponibilidad de datos en algunos de los municipios.

El mejor modelo estimado para estos últimos inmuebles, tras eliminar los valores atípicos, contiene como variables explicativas el tamaño de la nave comercial en metros cuadrados, la distancia al centro de la ciudad y la pertenencia al municipio de Dos Hermanas, y es el siguiente:

$$\widehat{Precio} = 161.996{,}672 + 378{,}836 \cdot Tamaño - 11{,}758 \cdot Dist. + 16.433{,}419 \cdot DosHerm.$$

A partir de esto podemos afirmar que a partir de un precio base de 161.996,67 €, el precio del inmueble aumenta en 378,84 € por cada metro cuadrado que aumente su tamaño, y disminuye a razón de 11,76 € por cada metro que se distancie del centro de la

ciudad. Además, si éste se encuentra situado en el municipio de Dos Hermanas, su precio aumenta en 16.433,42 €.

El coeficiente de determinación es de 0,802, por lo que el modelo explica más de un 80 por ciento de la variabilidad en el precio de una nave industrial de los municipios cercanos a la ciudad de Sevilla. Se ha verificado el cumplimiento de las hipótesis a priori para la construcción del modelo de regresión por el método de mínimos cuadrados.

Utilizando este modelo, una nave industrial situada en Dos Hermanas, con un tamaño medio de 378,84 metros cuadrados y a una distancia de 9.027 metros del centro, se estima que tendrá un precio de 384.365 €. Estas mismas condiciones, para una nave que no está situada en Dos Hermanas, tendría un precio estimado de 367.932 €.

Para la estimación del precio de una nave industrial de la ciudad de Sevilla, se ha utilizado el tamaño del inmueble, la disposición de aire acondicionado, si está situado en el distrito de Bellavista, o si lo está en Sevilla Este. El modelo estimado se muestra a continuación:

$$\widehat{Precio} = 528{,}50 \cdot Tam. + 352.967{,}19 \cdot Bellav. + 289.912{,}53 \cdot SevEste + 92.788{,}50 \cdot A.A.$$

A través de la estimación obtenida, el precio por metro cuadrado de las naves industriales de la capital es de 528,50 €, que para un tamaño medio de 690,64 metros cuadrados nos indica un precio medio de 365.000 €. Este valor aumenta en 289.913 € en el distrito de Sevilla Este, y prácticamente se duplica en Bellavista, alcanzando los 717.970 €. Además, la disponibilidad de aire acondicionado aumenta el valor en 92.788 €.

Las hipótesis requeridas para la estimación del modelo se verifican, y la variabilidad en el precio se explica en un 82,6% a través de éste.

Las diferentes características que determinan el precio de los locales comerciales y las naves industriales, hacen necesaria una separación clara en el estudio de estos inmuebles. La ciudad de Sevilla dispone de una amplia oferta de inmuebles comerciales disponibles a la venta, que es muy heterogénea en precio, tamaño y localización geográfica.

Los locales son inmuebles de reducida dimensión para los que lo más influyente a la hora de determinar su precio es, precisamente, su tamaño, así como la cercanía a zonas comerciales con gran tránsito de personas, tanto en el área urbana como en los municipios colindantes, donde su precio disminuye significativamente debido a una menor densidad de población y, por tanto, de actividad económica.

Las naves, por el contrario, se sitúan, en general, en zonas alejadas del centro de la ciudad donde el enlace con otras zonas de la región es de buena calidad. Igual que ocurre entre locales comerciales y naves industriales, las características que determinan el precio de una nave industrial dependen de si ésta se encuentra situada en la zona urbana de la capital o, por el contrario, en zonas periféricas.

Además, podemos afirmar que el precio de las naves industriales situadas en el municipio de Sevilla depende, en mayor intensidad, de su tamaño, y del hecho de tener aire acondicionado, lo que puede ser un indicativo de unas calidades superiores. También la ubicación en el distrito Sevilla Este y, en mayor medida el de Bellavista, suponen un incremento en el precio de la nave considerable.

Por otro lado, se ha evidenciado que el precio de las naves industriales del extrarradio mantiene una relación inversa con la distancia de las mismas al centro-ciudad. Por el contrario, la relación es directa entre el precio y el tamaño de la nave y, además, el precio aumenta en 16.000€ si se encuentra situada en el municipio de Dos Hermanas.

Al comparar los resultados obtenidos para el modelo de estimación del precio de los locales comerciales con el estudio realizado por Rey (2014), realizado en la ciudad de Córdoba, observamos que el tamaño y la ubicación son relevantes en ambos modelos.

Las características que determinan el precio de un local comercial y su relación de influencia con éste, difieren de las observadas para las naves industriales. Además, la relación entre las variables explicativas y el precio varía en función de la región en la que se realiza el análisis, ya que las características propias del entorno pueden modificar la valoración que cada una de las características tienen.

Como ejemplo, cabe destacar la importancia que la presencia de aire acondicionado tiene en la valoración de los locales comerciales de la ciudad de Sevilla, motivado por las elevadas temperaturas que en esta ciudad se alcanzan en períodos estivales.

Por otro lado, las naves industriales son inmuebles en los que el tamaño y la ubicación son las únicas variables relevantes en su valoración.

Por todo ello, para alcanzar un profundo conocimiento del mercado inmobiliario de un país, el análisis debe realizarse a nivel local y para las distintas tipologías de inmueble, ya que la gran heterogeneidad de este mercado hace que sus características tengan una influencia significativamente diferente sobre su valoración.

9.Conclusiones

En este último capítulo se resumirán, a modo de conclusión, los objetivos alcanzados, exponiendo las dificultades halladas, así como las limitaciones de las que el autor es consciente. Del mismo modo, se propondrán líneas de investigación y colaboraciones, empresa – Universidad, futuras que pueden generarse a partir del trabajo realizado, algunas de las cuales se encuentra ya en marcha.

Es un trabajo multidisciplinar que ha exigido elevados conocimientos de métodos estadísticos, econométricos, de desarrollo de software, así como, de los relativos al mercado inmobiliario, sector objeto y beneficiario de los resultados obtenidos.

El mercado inmobiliario, y en particular el español, se caracteriza por su elevado dinamismo, su alta heterogeneidad y un gran volumen de negocio. Esto, obliga a disponer de una gran cantidad de información que sea fiable, inmediata y fácil de analizar, para poder llevar a cabo un correcto estudio de sus características.

El objetivo principal, planteado al inicio de la investigación, es dar respuesta a la necesidad de conocer el mercado profundidad, expresada en numerosas ocasiones por los distintos agentes, y acrecentada por la reciente crisis inmobiliaria vivida en el país, aún en proceso de lenta recuperación. Para ello, se ha diseñado una completa metodología que

podríamos denominar *llave en mano*, ya que permite una ágil valoración de distintos inmuebles, a la venta o en alquiler, de una determinada región.

Los pasos seguidos para la construcción de esta metodología son:

- Búsqueda de una fuente fiable de información que contenga, además, una cantidad de datos que permita un fiable y robusto análisis.
- Diseño de un sistema automatizado de recogida de la información disponible en las fuentes halladas.
- Análisis y discriminación de las metodologías de valoración a aplicar en función de la información disponible.
- Combinación de las distintas metodologías aplicables para la obtención de la valoración óptima.
- Implementación de dichas metodologías en un software de fácil manejo, capaz de trabajar con multitud de datos, de forma que pueda ser usado sin necesidad de conocimientos sobre programación.
- Aplicación del software en diferentes zonas geográficas para la estimación del precio de diferentes inmuebles, que permita, además, la evaluación de la potencia del mismo.

El método de valoración por comparación, y en concreto, el método econométrico o de regresión es, finalmente, el propuesto en este trabajo, conclusión a la que se ha llegado tras una detallada revisión de los métodos de valoración existentes, y debido a la elevada capacidad predictiva de las estimaciones realizadas, que se llevan a cabo a partir de un mayor número de variables explicativas.

Fuentes de información

Para la búsqueda de fuentes proveedoras de información, se ha optado por los portales web inmobiliarios con servicio a nivel nacional, ya que es la fuente más rica de información, que además se caracteriza por ser de ámbito español, experimentar una continuas y rápidas actualizaciones, y ser la de más fácil acceso para el investigador. Otro punto a favor de esta elección es la rápida integración en el mercado que éstos han tenido.

Prácticamente, desde que en el año 2000 proliferaran numerosos portales en este sector, y que han ido mejorando en prestaciones y recursos ofertados con los años, hay un portal que ha destacado sobre los demás, por su gran oferta de anuncios y su elevado tráfico web. Este es idealista.com.

A cierta distancia encontramos fotocasa.es que con unos números muy importantes sigue los pasos del portal líder. En la continua renovación que se produce de la oferta de portales inmobiliarios, en la actualidad destacan también pisos.com y yaencontre.com.

En los últimos años, también han proliferado los denominados portales agregadores, que reúnen la oferta de distintos portales inmobiliarios con bases de datos propias. Estos portales están apoyados, en su mayoría, por pisos.com que autoriza el enlace a sus anuncios, lo que le reporta a su vez un mayor tráfico web, ya que los usuarios de los agregadores son finalmente redirigidos a su portal.

El inconveniente principal de los portales agregadores es la falta de efectividad en las búsquedas en las que aparecen tipos de anuncios de inmuebles no solicitados, anuncios duplicados o incluso triplicados y, en ocasiones, ofertas de provincias distintas a la solicitada. A esto hay que añadir, que el tráfico que concentran es todavía muy inferior al de los principales portales inmobiliarios con oferta propia. Por otro lado, la gran oferta de estos portales, hace de ellos una fuente de información a considerar.

De todo ello, se ha concluido que el portal gestionado por la empresa Idealista, S.A. reúne las mejores condiciones para ser utilizado como fuente de información de estudios relativos al mercado inmobiliario de oferta en España. Su elevada oferta inmobiliaria y su extensa implantación a nivel nacional como medio de difusión de inmuebles, tanto por parte de particulares como de empresas del sector, hacen de idealista un importante proveedor de datos para el análisis.

Existe, además, con este portal, una oportunidad real de ampliación del marco geográfico de análisis, debido a su progresiva implantación en otros países como Italia y Portugal.

Software desarrollado

Un acuerdo de colaboración con la empresa propietaria del portal Idealista.com, ha permitido el acceso, con las lógicas restricciones impuestas para la protección de la información, a sus datos; a través de una aplicación desarrollada por completo en este trabajo y que integra diferentes tecnologías de programación como JAVA, JSON, H2, Eclipse Link, ControlsFX o jsoup.

En él se han integrado numerosas librerías desarrolladas previamente con otras nuevas desarrolladas en este trabajo, de forma que da resultado es un programa informático multiplataforma, que permite la automática recogida de los datos de una región, la manipulación de los datos obtenidos y su posterior procesamiento estadístico. En la elección de las librerías utilizadas ha sido determinante el hecho de que todas ellas son de libre distribución.

Todo el entorno gráfico ha sido diseñado previamente de forma que resultara sencillo e intuitivo. No obstante, debido a las funcionalidades estadísticas incluidas en el mismo, son necesarios conocimientos de la materia para su correcta utilización.

Una integración entre el software implementado, desarrollado en Java, y el motor del software estadístico R, a través de librerías como RJava y JRI, ha permitido desarrollar una herramienta que permite realizar todo el proceso, desde la recogida de datos hasta el análisis estadístico de los mismos, incorporando, además, la construcción de modelos de regresión y de redes neuronales con variables cualitativas, su validación y la detección de observaciones que pudieran influir en exceso en la construcción de dichos modelos.

Se ha diseñado un proceso de adquisición de los datos del portal que permite definir proyectos de cuatro tipos de inmuebles: viviendas, locales comerciales, oficinas y garajes, fijando el centro de búsqueda y un radio máximo de 15 kilómetros. Mediante un acceso autorizado al portal, se recoge dicha información, se realiza la conversión a variables manipulables por el software, y se almacenan en la base de datos para su posterior consulta, filtrado o análisis. Las características de los inmuebles recogidas en esta búsqueda son de tres tipos: las propias de la localización del inmueble, las del propio anuncio y las relativas a las condiciones del inmueble.

Del mismo modo, ha sido implementada una posible actualización de los datos en el tiempo que permita analizar la evolución del mercado, en relación al cambio en la oferta o a la variación del precio de venta o alquiler de los distintos inmuebles.

Como se ha mencionado anteriormente, y tras una extensa revisión bibliográfica, se han implementado modelos de estimación de precios, basados en regresión y en redes neuronales, junto a los procedimientos de validación necesarios.

En resumen, se ha creado un programa cuyas funcionalidades son: recogida de los datos relativos a una vivienda, local, oficina o garaje, a la venta o en alquiler, de cualquier punto de la geografía nacional, en zonas circulares de hasta 15 kilómetros de radio, creación de una base de datos con esta información, actualización en el tiempo, manipulación de los datos como ordenación o filtrado de los mismos, exportación de los datos a formato CSV para ser analizados con otro programa estadístico, análisis estadístico de estos datos con el mismo programa, a través de su enlace con el software estadístico R, obtención en formato html de los resultados obtenidos y guardado o impresión de los mismos.

Desde el punto de vista de la estadística, el programa permite la realización de un estudio descriptivo de los datos, tanto univariante como bivariante, pruebas de normalidad, intervalos de confianza, estimación de modelos de regresión, no sólo para el precio, lineales o no lineales, que puede incluir variables cualitativas a través de la introducción de variables artificiales, realización de un completo proceso de validación del modelo estimado, estimación a través una red de tipo perceptrón multicapa con diferentes configuraciones, detección de observaciones influyentes que puedan desvirtuar el modelo estimado y elaboración de mapas de información geográfica de los datos.

Observaciones influyentes

La gran influencia que sobre el modelo de estimación tienen los inmuebles con características extraordinarias, hace necesario el estudio de la presencia, en la muestra, de dichas observaciones para la posterior toma de decisión relativa a su validación.

Tras comprobar el errático funcionamiento que en ocasiones tienen los métodos clásicos, por lo se propone un método para la detección de observaciones influyentes basado en el cambio del valor del criterio de información del modelo de regresión construido para la estimación de valores como el precio. Para evaluar el funcionamiento de éste y del resto de métodos, se ha aplicado un proceso de Monte Carlo que ha reunido cientos de miles de simulaciones. El método propuesto ha revelado ser, en muchos casos, mejor que los métodos clásicos.

El desconocimiento de la distribución del criterio de información y, por consiguiente, la del cambio producida en ésta en presencia de una observación influyente, hace

imposible la demostración teórica de la idoneidad de un método, basado en el análisis del cambio producido en el criterio de información al eliminar esta observación del conjunto de datos.

Se ha comprobado, mediante estas simulaciones, que la proporción de cambio en el criterio de información producida al eliminar una observación del conjunto de datos, es significativamente superior si la observación eliminada es influyente.

Por todo ello, si el valor del criterio de información disminuye significativamente al eliminar una observación, ésta será influyente. De este modo, para un conjunto de n observaciones, si se construyen n modelos de regresión, para los conjuntos de n-1 datos resultantes de eliminar una observación cada vez, y para todos ellos el valor de su criterio de información, el vector de criterios de información obtenido nos permitirá detectar la presencia de observaciones influyentes.

Mercado de la oferta de viviendas a la venta en Sevilla

Para mostrar la potencia de los métodos de valoración integrados, se han llevado a cabo dos análisis de mercado, el de viviendas a la venta de la ciudad de Sevilla, y el de los locales comerciales y las naves industriales de la aglomeración urbana de Sevilla que han arrojado los resultados que se muestran a continuación.

Para el primero de los análisis se ha utilizado una muestra formada por más de 10000 inmuebles, con más de veinte variables descriptivas, entre las que se encuentra el precio de venta de la vivienda, cuya media está situada en unos 274363 €, destacando los barrios de los Remedios, La Palmera, Centro, La Palmera, Heliópolis, Santa Clara y Nervión.

Por otro lado, el tamaño medio de este tipo de inmuebles en la ciudad de Sevilla, es de 132 m^2. Combinando ambas características, se estima que el precio medio por metro cuadrado de la ciudad de Sevilla asciende a 1971 €, por encima del registrado a nivel nacional a principios de 2017, situado en algo más de 1636 €.

Las viviendas de Sevilla disponen de 3,25 habitaciones y 1,82 baños, por término medio. Son mayoritariamente pisos, sin cochera ni ascensor ni piscina, que tienen aire acondicionado ni armario, y que tiene un estado bueno de conservación.

También se han construido, a nivel general y en algunos distritos, modelos de estimación de precio mediante la metodología de precios hedónicos y la construcción de redes neuronales de tipo perceptrón multicapa, con buenos resultados. Previamente, han sido identificado los inmuebles influyentes en la construcción.

Mercado de la oferta de inmuebles comerciales a la venta en la aglomeración urbana

En el segundo caso práctico de aplicación, se han analizado los locales comerciales, son aquellos inmuebles en los que se lleva a cabo algún tipo de actividad económica. No obstante, las diferentes características que determinan su precio, hacen necesario un estudio diferenciado entre locales comerciales y naves industriales. La estimación del precio a través de la metodología de precios hedónicos es inédita en el caso de las naves industriales y sumamente escasa para los locales comerciales, al contrario de lo que ocurre con los inmuebles destinados a vivienda.

Como hemos visto, el precio de una nave industrial de la periferia de la ciudad de Sevilla, está íntimamente relacionado con su tamaño y con la distancia al centro de la ciudad, de forma que la valoración estimada del metro cuadrado se sitúa en torno a los 379 €, y por cada kilómetro que aumente la distancia, el precio se devalúa en casi 12000 €. Además, en el municipio de Dos Hermanas el precio es unos 16000 € superior al resto de municipios colindantes.

Al contrario que ocurre con las naves industriales, la valoración del precio por metro cuadrado de un local comercial, disminuye cuando el tamaño de éste aumenta de forma ligeramente inelástica, de modo que un aumento de 1 % del tamaño de un local, produce un incremento del precio del 0,913 %. Otro factor que produce un incremento en el precio, aunque en menor medida, es la cercanía al centro de la ciudad, donde el precio por metro cuadrado es el más elevado de toda la región, a gran diferencia del resto.

Debido a las elevadas temperaturas alcanzadas por la ciudad en verano, el aire acondicionado es un elemento que se hace indispensable en un local comercial. Esto se refleja en un incremento en la valoración del precio de éste, superior al 12 % del valor total, cuando el inmueble dispone de climatización, lo que supone que para el precio medio de un local que como calculamos anteriormente era de 267.331,26 €, el incremento en su valoración cuando dispone de aire acondicionado es superior a los 32000 €.

Limitaciones encontradas

En la elaboración de este trabajo se han detectado algunas limitaciones, algunas de las cuales requieren de continuidad en nuestro trabajo y otras de agentes externos. Fundamentalmente, distinguimos tres tipos, las relativas a los datos disponibles, a la implementación del software, y al método de detección de observaciones influyentes desarrollado. Se detallan a continuación.

Limitaciones en los datos: No es posible disponer de nuevos campos, que pudieran influir significativamente en el proceso de estimación, como puede ser la antigüedad del edificio. Los continuos contactos con la empresa son esenciales para el progresivo incremento de la información disponible, tanto para la investigación como para el funcionamiento de la empresa.

Limitaciones del software: La restricción del número mensual de peticiones de información y el máximo radio de búsqueda, son limitaciones importantes de la metodología de recogida de datos, así como la posibilidad de que la información recogida del servidor sea incorrecta o inexacta. El sistema de introducción de datos de la empresa en el servidor hace que está posibilidad se reduzca drásticamente. Por otro lado, la necesidad de una contraseña proporcionada por la empresa para la realización de peticiones en el servidor. Esta restricción está justificada en la necesidad de la empresa en proteger sus datos. El acceso libre al mismo para fines académicos incrementaría el conocimiento del mercado inmobiliario español.

Método de detección de observaciones influyentes: Se ha analizado la bondad del método, sólo para el caso bivariante, por lo que es necesario continuar el trabajo en la línea de evaluar cómo influye en el resultado de las simulaciones, la inclusión de más variables explicativas.

Líneas de investigación abiertas

Las líneas que deja abiertas este trabajo son las siguientes:

- Incremento del mercado geográfico de las búsquedas, incluyendo países como Italia y Portugal.
- Actualización del sistema de búsqueda al nuevo sistema lanzado por la empresa recientemente, que incorpora nuevos campos con características del inmueble.

- Optimización del proceso de consulta de la base de datos actual.
- Análisis del método de detección de observaciones influyentes en modelos de regresión múltiple.
- Incorporación de nuevos métodos para la estimación, como puede ser el análisis funcional o el método de mínimos cuadrados parciales.
- Incorporación de información macroeconómica al análisis, que aumente la riqueza de la información.

Por todo lo analizado, podemos dar por logrados los objetivos planteados inicialmente en este trabajo.

Concluimos, de todo ello, que las líneas de investigación desarrolladas en este trabajo, relativas al mercado inmobiliario, pueden aportar nuevos enfoques para su análisis, tanto en los ámbitos macroeconómicos como microeconómicos.

Los resultados de aplicación derivados de este trabajo permitirán a actores del mercado: arrendatarios, propietarios particulares, inversores profesionales, agencias inmobiliarias, promotores y administración pública; disponer de mayor información para la toma de decisiones, mayor transparencia y conseguir mejoras en la productividad, y la rentabilidad de sus actividades.

Conclusions

In this last chapter we will summarize, by way of conclusion, the objectives reached explaining the difficulties encountered, as well as the limitations of which the author is aware. In the same way, future research lines and collaborations, company - University, will be proposed that can be generated from the done work.

It is a multidisciplinary work that has required high knowledge of statistical methods, econometrics, software development, as well as those related to the real estate market, object sector and beneficiary of the results.

The real estate market, and in particular Spanish, is characterized by its high dynamism, its high heterogeneity and a large volume of business. This requires a large amount of information that is reliable, immediate and easy to analyze, in order to carry out a proper study of its characteristics.

The main objective, raised at the beginning of the investigation, is to respond to the need to know the market in depth, expressed on numerous occasions by the different agents, and increased by the recent real estate crisis experienced in the country, still in the process of slow recovery . For this purpose, a complete turnkey methodology has been designed, which allows a quick valuation of the properties of a region.

The steps followed for the construction of this methodology are:

- Search for information sources
- Design of an automated system for collecting information from the found sources
- Analysis and discrimination of valuation methodologies to be applied based on available information
- Combination of the different methodologies applicable to obtain the optimal valuation.
- Implementation of these methodologies in an easy-to-use software capable of working with a multitude of data.
- Application of the software in different geographical areas for the evaluation of its power.

The method of measurement by comparison and in particular, the econometric or regression method, is finally the one proposed in this work, conclusion reached after a detailed review of the existing valuation methods and due to the high predictive ability of the estimates made, which are carried out from a greater number of explanatory variables.

Sources providing information

For the search of sources providing information we have opted for real estate web portals with service at the national level, being the richest source of information of Spanish scope with a quick update and of easier access for the researcher . Another point in favor of this election is the fast integration in the market that they have had. The results have been published in Casas, Caridad and Núñez (2017)

Practically, numerous portals proliferated in this sector since the year 2000 and have been improving in services and resources offered over the years, there is a portal that has stood out above the others due to its large offer of ads and its high web traffic . This is idealista.com.

At some distance we find fotocasa.es that with very important numbers follows the steps of the leading portal. In the continuous offer of real estate portals renovation pisos.com and yaencontre.com highlights nowadays.

In recent years, there have also proliferated the so-called aggregation portals, which bring together the offer of different real estate portals with their own databases. These portals are mostly supported by pisos.com that authorizes the link to their ads, which in

turn reports a greater web traffic, since the users of the aggregators are finally redirected to their portal.

The main drawback of aggregator portals is the lack of effectiveness of searches in which types of advertisements for unsolicited properties appear, duplicate or even triplicate ads, and sometimes offers from provinces other than the one requested. To this, it must be added that the traffic they concentrate is still much lower than that of the main real estate portals with their own offer. On the other hand, the great offer of these portals makes them a source of information to consider.

From all this, it has been concluded that the portal managed by the company Idealista, S.A. has the best conditions to be used as a source of information on studies related to the real estate offer market in Spain. Its high real estate offer and its extensive implementation at national level as a means of spreading real estate, both by individuals and companies in the sector, make idealista a major data provider for analysis.

There is also a real opportunity to expand the geographical framework of analysis, due to its progressive implementation in other countries such as Italy and Portugal.

Software

A collaboration agreement with the company that owns the Idealista.com portal has allowed access, with the logical restrictions imposed for the protection of information, to its data; through an application completely developed in this work and that integrates different programming technologies like JAVA, JSON, H2, Eclipse Link, ControlsFX or jsoup.

A portal data acquisition process has been designed to define four types of real estate projects: homes, commercial premises, offices and garages, setting the search center and a maximum radius of 15 kilometers. Through an authorized access to the portal, this information is collected, the conversion is made to variables manipulated by the software and stored in the database for later consultation, filtering or analysis. The characteristics of the properties collected in this search are of three types: those of the location of the property, those of the advertisement itself and those relating to the conditions of the property.

In the same way, a possible update of the data has been implemented in the time that allows analyzing the evolution of the market in relation to the change in the offer or to the variation of the sale or rental price of the different properties.

As mentioned above and following an extensive literature review, regression-based pricing models and neural networks have been implemented along with the necessary validation procedures.

Influential observations

The great influence that the properties have with extraordinary characteristics on the estimation model makes it necessary to study the presence in the sample of these observations for the subsequent decision on their validation.

After verifying the erratic operation that sometimes have the classic methods, a method is proposed for the detection of influential observations based on the change of the value of the criterion of information of the constructed regression model for the estimation of values like price. To evaluate the performance of this and other methods, a Monte Carlo process has been applied and it has gathered hundreds of thousands of simulations. The proposed method has proved to be, in many cases, better than the classical methods.

An integration between the implemented software, developed in Java, and the R software engine, through libraries such as RJava and JRI, has allowed to develop a tool that allows to carry out the whole process, from the data collection to its statistical analysis incorporating, in addition, the construction of regression models and neural networks with qualitative variables, their validation and the detection of observations that could influence in the construction of such models in excess.

Therefore, once the data has been collected, the software developed allows a previous descriptive analysis to be carried out, including a univariate, bivariate and geographical study, since the program has the ability to elaborate geographic information systems. Next, the variables to be used in the model are decided and for these are detected and eliminated, if necessary, any influential observations. Finally, the regression models and neural networks are constructed and validated and the results obtained can be compared.

Seville: Two market analyzes

In order to show the strength of the integrated valuation methods, two market analyzes have been carried out, the houses for sale in the city of Seville and the commercial premises and industrial buildings in the urban agglomeration of Seville that have given the results shown below.

For the first analyzes, a sample of more than 10000 buildings was used, with more than twenty descriptive variables among which the sale price of the house is, whose average is situated at about 274363 €, highlighting the neighborhoods Los Remedios, La Palmera, Centro, La Palmera, Heliopolis, Santa Clara and Nervión.

On the other hand, the average size of this type of real estate in the city of Seville is 132 m^2. Combining both characteristics, it is estimated that the average price per square meter of the city of Seville amounts to € 1971, above the national level at the beginning of 2017 situated at just over € 1636.

Houses of Seville have 3.25 rooms and 1.82 bathrooms, by average term. They are mainly flats, without garage, elevator or swimming pool, which have air conditioning or closet, and that have a good state of conservation.

Models of estimation price using the hedonic price methodology and the construction of neuronal networks of multilayered percepetron type have also been constructed, in general and in some districts, with good results. Previously, influential buildings have been identified in the construction.

In the second practical application case, the commercial premises have been analyzed, they are those buildings in which some type of economic activity is carried out. However, the different characteristics that determine their price, make it necessary to study differently between commercial premises and industrial warehouses. The estimation of the price through the methodology of hedonic prices is unprecedented in the case of industrial warehouses and extremely scarce for commercial premises, unlike what happens with real estate intended for housing.

As we have seen, the price of an industrial warehouse on the outskirts of the city of Seville is intimately related to its size and distance to the city center, so that the estimated

value of the square meter is around the 379 €, and for each kilometer that increases the distance the price devalues by almost 12000 €. In addition, in the municipality of Dos Hermanas the price is about € 16000 higher than the rest of neighboring municipalities.

In contrast to industrial warehouses the price per square meter of a commercial space decreases when the size of the warehouse increases slightly inelastically, so that an increase of 1% in the size of a room produces an increase of the price of 0.913%. Another factor that produces an increase in price, although to a lesser extent, is the proximity to the center of the city, where the price per square meter is the highest in the whole region, unlike the rest.

Due to the high temperatures reached by the city in summer, air conditioning is an element that becomes indispensable in a commercial place. This is reflected in an increase in its valuation of the price, more than 12% of the total value, when the property has air conditioning, which means that for the average price of a premises that as we previously estimated was 267,331.26 €, the increase in its valuation when it has air conditioning is higher than 32000 €.

Limitations

Finally, we will analyze the limitations of this work.

Limitations on data: It is not possible to have new fields that could significantly influence the estimation process. The continuous contacts with the company are essential for the progressive increase of the available information, both for the investigation and for the operation of the company.

Limitations of the software: The restriction of the monthly number of requests for information and the maximum search radius are important limitations of the data collection methodology, as well as the possibility that the information collected from the server is incorrect or inaccurate. The data entry system of the company on the server makes this possibility drastically reduced. On the other hand, the need for a password provided by the company to make requests on the server. This restriction is justified by the company's need to protect its data. Free access to it for academic purposes would increase knowledge of the Spanish real estate market.

Method of detection of influential observations: We have analyzed the goodness of the method, only for the bivariate case, so it is necessary to continue the work in the line of evaluating how it influences the result of simulations, the inclusion of more explanatory variables.

Future lines of research

The future lines of research that leave this work open are the following:

- Increase in the geographic market for searches, including countries like Italy and Portugal.
- Updating the search system to the new system launched by the company recently, which incorporates new fields with characteristics of the property.
- Optimization of the query process of the current database.
- Analysis of the method to detect influential observations in multiple regression models.
- Incorporation of new methods for estimation, such as functional analysis or partial least squares method.
- Incorporation of macroeconomic information to the analysis, which increases the wealth of information.

For all this, we can grant as achieved the objectives initially set out in this work.

From all this it can be concluded that the lines of research developed in this work related to the real estate market can provide new approaches for the analysis of the market both in the macroeconomic and microeconomic fields.

The application results derived from this work will allow market players: tenants, private owners, professional investors, real estate agencies, promoters and public administration; to have more information for decision-making, greater transparency and achieve improvements in productivity and profitability of its activities.

10. Bibliografía

Bibliografía

Akaike, H. (1977). On entropy maximization principle. *Application of statistics*, 543, 27-41.

Ayuntamiento de Sevilla. (2016). Portal de datos abiertos. Sevilla. Recuperado de: http://datosabiertos.sevilla.org/

Ballestero, E. (1973). Notas sobre un nuevo metodo rapido de valoración. Revista de estudios agro-sociales, 85, 75-78.

Bayer, P., Keohane, N., y Timmins, C. (2009). Migration and hedonic valuation: The case of air quality. *Journal of Environmental Economics and Management, 58*(1), 1-14.

Belsley, D. A., Kuh, E. y Welsch, R. E. (1980). Detecting and assessing collinearity. En John Wiley & sons, Inc. (Ed.), *Regression diagnostics: Identifying influential data and sources of collinearity*, pp. 85-191. E.E.U.U.

Breusch, T. y Godfrey, L. (1981). A review of recent work on testing for autocorrelation in dynamic linear models. *Macroeconomic Analysis: Essays in Macroeconomics and Macroeconometrics.* London: DA Currie

Breusch, T. y Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica, 47(5), 1287-1294. doi:10.2307/1911963

Caballer, V. y Mellado, V. C. (1998). Valoración agraria: teoría y práctica. Madrid: Mundi-Prensa Libros.

Caridad, J.M. y Brañas, P. (1996). "El mercado de la vivienda en Córdoba: un enfoque cuantitativo". XXII Reunión de Estudios Regionales. Pamplona.

Caridad, J. M. (1998). *Econometría: modelos econométricos y series temporales.* Barcelona: Reverté.

Caridad, J. M., Núñez, J. M., y Ceular, N. (2008). Metodología de precios hedónicos vs. Redes Neuronales Artificiales como alternativas a la valoración de inmuebles. Un caso real. *CT: Catastro*, 62, 27-42.

Caridad, J. M. y Ceular, N. (2001). Un análisis del mercado de la vivienda a través de redes neuronales artificiales. *Estudios de economía aplicada,* 18, 67-81.

Casas, J.C., Caridad y Ocerin, J. M., Núñez Tabales, J. (2016), Una visión del mercado inmobiliario digital. Comparativa de los principales portales inmobiliarios y agregadores de oferta española. *Ar@cne. Revista Electrónica de Recursos en Internet sobre Geografía y Ciencias Sociales,* 204

Chatterjee, S., y Hadi, A. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, 1(3), 379-393. Recuperado de: http://www.jstor.org/stable/2245477

Clarkson, D. B., et al. (1993). A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in r x c contingency tables. *ACM Trans. Math. Softw*. 19(4): 484-488.

Colwell, P. F. y Dilmore, G. (1999). Who Was First? An Examination of an Early Hedonic Study. *Land Economics, 75*(4), 620-626. doi: 10.2307/3147070

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics,* 19(1), 15-18.

Court, A. T. (1939). Hedonic Price Indexes with Automotive Examples. En General Motors Corporation (Ed.) *The Dynamics of Automobile Demand,* Detroit, Michigan.

Davidson, R., y MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society*, 49(3), 781-793.

Dios, R. (1998). El modelo lineal sin término independiente y el coeficiente de determinación. Un estudio Monte Carlo. *Qüestiió,* 22(1), 3-37.

Diputación Provincial De Sevilla (2016). Anuario Estadístico de la Provincia de Sevilla (20). Recuperado de

http://portalestadistico.dipusevilla.es/galeriaFicheros/anuario_estadistico/anuario2016.pdf

Durbin, J. (1970). Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables. *Econometrica*, 38(3), 410-421. Doi:10.2307/1909547

Durbin, J., y Watson, G. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, 37(3/4), 409-428. doi:10.2307/2332391.

Ecma International (2016). The API platform for digital business: Apigee. Recuperado de: https://apigee.com/api-management/#/homepage

Facilísimo Interactive (2015). Pisos en Madrid, pisos en Barcelona, pisos en Valencia… Madrid: Expocasa.com. Recuperado de: http://www.expocasa.com/

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* Boca Ratón: Taylor and Francis Grooup.

Fisher, R. (1935). The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, 98(1), 39-82. doi:10.2307/2342435

Fomento (2016). Información para el ciudadano. Información estadística del ministerio de Fomento. Recuperado de: https://www.fomento.gob.es/MFOM/LANG_CASTELLANO/ATENCION_CIUDADANO/INFORMACION_ESTADISTICA/

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ... y Heiberger, R. (2016). Package 'car'. Companion to Applied Regression. R Package version, 2-1. Recuperado de: ftp://mirrors.ucr.ac.cr/CRAN/web/packages/car/car.pdf

Fuerst, F. y McAllister, P. (2011), Green Noise or Green Value? Measuring the Effects of Environmental Certification on Office Values. *Real Estate Economics*, 39: 45–69. doi:10.1111/j.1540-6229.2010.00286.x.

Garson, D. G. (1991). Interpreting neural network connection weights. *AI Expert*, 6(7), 47-51

Gibbons, S. y Machin, S. (2008). Valuing school quality, better transport, and lower crime: evidence from house prices. *Oxford review of Economic Policy, 24*(1), 99-119.

Giles, J. (2016). ControlsFX, JavaFX News, Demos and Insight: FX Experience. Recuperado de: http://fxexperience.com/controlsfx/

Globaliza (2015). Casas en venta y alquiler en España. Madrid: Globaliza.com. Recuperado de: https://www.globaliza.com/

Goldfeld, S. M. y Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American statistical Association, 60*(310), 539-547.

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric of quality change En NBER (Ed.) , *The Price Statistics of the Federal Goverment* (pp. 173-196). Price Statistics Review Comitee.

Gross, J. y Ligges, U. (2012). nortest: Tests for Normality. R package version 1.0. Recuperado de: http://CRAN. R-project. org/package= nortest.

Guadalajara N. (2014). *Métodos de valoración inmobiliaria*. Madrid: Ediciones Mundi-Prensa.

Gujarati, D. N., y Porter, D. C. (2011). *Econometria Básica-5*: Bogotá: McGraw Hill.

Günther, F. y Fritsch, S. (2010). Neuralnet: Training of neural networks. *The R journal, 2*(1), 30-38. Recuperado de: https://datajobs.com/data-science-repo/Neural-Net-[Gunther-and-Fritsch].pdf

HabitatSoft (2015). Casas y pisos en venta, alquila o vende tu piso - Pisos.com. Madrid: Pisos.com. Recuperado de: https://www.pisos.com/

Hass, G. (1922). Sale Prices as Basis for Farm Land Appraisal. The University of Minnesota Agricultural Experiment Station. *Technical Bulletin*, 9.

Haykin, S. (2001). *Neural Networks: A Comprehensive Foundation*. Tsinghua University Press.

Hebb, D. O. (1949). The organization of behavior: A neuropsychological approach. New York: John Wiley & Sons.

Hecht-Nielsen R. (1989) Neurocomputer Applications. En: Eckmiller R., v.d. Malsburg C. (Eds.) *Neural Computers*. Springer Study Edition, vol 41. Springer, Berlin, Heidelberg

Hendry, D. F. (1984). Monte Carlo experimentation in econometrics. Handbook of econometrics, 2, 937-976. doi: https://doi.org/10.1016/S1573-4412(84)02008-0

Hoaglin, D. C. y Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1), 17-22.

Hoerl, A. E. y Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hogaria.net (2015). Pisos en Venta y Alquiler, Comprar Casa, Pisos Baratos - Hogaria. Madrid. Recuperado de: https://www.hogaria.net/

Hope, A. C. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society*. Series B (Methodological), 30(3), 582-598. Recuperado de: http://www.jstor.org/stable/2984263

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., Mitchell, D., y Zeileis, M. A. (2015). Package 'lmtest': Testing Linear Regression Models. Recuperado de: https://cran.r-project.org/web/packages/lmtest/lmtest.pdf

Iberanuncios (2015). Fotocasa.es: Alquiler de pisos, Venta de pisos en Madrid, Barcelona, Valencia… - Tucasa.com. Recuperado de: https://www.tucasa.com

Idealista (2015). Casas y pisos, alquiler y venta. Anuncios gratis. Madrid: Idealista.com. Recuperado de: https://www.idealista.com/

Instituto Nacional De Estadística (2016). INEbase. Madrid: INE. Disponible en: http://www.ine.es/

Jarque, C. M. y Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255-259.

Kahle, D. y Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R journal*, *5*(1), 144 - 161. Recuperado de: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Karagrigoriou, A., Mattheou, K. y Vonta, I. (2011). On asymptotic properties of AIC variants with applications. *Open Journal of Statistics*, 1(2), 105-109.

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for statistical applications and methods, 22*(6), 665-674.

Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics*, 21(2), 193-199.

Kitagawa, G., & Akaike, H. (1982). A quasi Bayesian approach to outlier detection. *Annals of the Institute of Statistical Mathematics, 34(1),* 389-398.

Kleinbaum, D., Kupper, L., Nizam, A. y Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods*. Nelson Education.

Kohonen (1988). An introduction to neural computing. *Neural networks*, 1(1), 3-16.

Kolmogorov, A. H. (1933). Sulla determinazione empirica di una leggi di distribuzione. Giorn. 1st it lit o Ital. Attuari, 4, 83-91.

Kong, F., Yin, H., y Nakagoshi, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*, *79*(3), 240-252.

Kornacki, A., Kyureghyan, K., y Ignaciuk, S. (2012). Application of Akaike information criterion for the detection of outliers. *Teka Komisji Motoryzacji i Energetyki Rolnictwa*, 12(2), 111-115

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of political economy*, *74*(2), 132-157.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, *62*(318), 399-402.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68-78.

McCarthy, J. (1956, Julio). "Measures of the value of information". Comunicación presentada en Dartmouth College, Hanover, New Hampshire.

McClelland, J. L., Rumelhart, D. E., y PDP Research Group. (1987). Parallel distributed processing (Vol. 2). Cambridge, MA: MIT press.

McCulloch, W. S. y W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4): 115-133.

McLeod, A. I. (2005). Kendall rank correlation and Mann-Kendall trend test. *R Package "Kendall"*. Recuperado de: https://cran.r-project.org/web/packages/Kendall/Kendall.pdf

Mehta, C. R. y Patel, N. R. (1986). Algorithm 643: Fexact: A Fortran subroutine for Fisher's exact test on unordered r× c contingency tables. *ACM Transactions on Mathematical Software (TOMS)*, *12*(2), 154-161.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. y Lin, C. (2015). Misc Functions of the Department of Statistics. *Probability Theory Group (Formerly: E1071), TU Wien*. Recuperado de: https://cran.r-project.org/web/packages/e1071/e1071.pdf

Minsky, M. and Papert S. (1969). *Perceptrons. An introduction to computational geometry (1st ed.)*. MA.: MIT Press, CAmbridge.

Mitula Group (2015). Buscador imobiliario con pisos en alquiler y venta - Mitula Pisos. Madrid: Pisos.mitula.com. Recuperado de: https://pisos.mitula.com/

Mitula Group (2015). Pisos y casas en venta y en alquiler en toda España - Nestoria. Madrid: Nestoria.es. Recuperado de: https://www.nestoria.es/

Mitula Group (2015). ¡Encuentra tu casa! - Nuroa.es. Madrid: Nuroa.es. Recuperado de: https://www.nuroa.es/

Montaño, J. J., Palmer, A. y Fernández, C. (2002), Redes neuronales artificiales: abriendo la caja negra, *Metodología de las ciencias del comportamiento* 4(1): 77-93.

Montgomery, D. C., Peck, E. A., y Vining, G. G. (2015). *Introduction to linear regression analysis*. Hoboken, New Jersey: John Wiley & Sons.

Murata, N., Yoshizawa, S. y Amari, S. I. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks* 5(6): 865-872.

Núñez, J. M. (2007). *Mercados Inmobiliarios: Modelización de los Precios.* (Tesis Doctoral). Universidad de Córdoba, España

Oracle And/Or Its Affiliates (2015). Java Platform, Standard Edition. Deployment Guide. Capítulo 10: JavaFX Ant Tasks. Recuperado de: http://docs.oracle.com/javase/8/docs/technotes/guides/deploy/

Pope, J. C. (2008). Buyer information and the hedonic: the impact of a seller disclosure on the implicit price for airport noise. *Journal of Urban Economics*, *63*(2), 498-516.

Quesada, F. J. G., Graciani, M. A. F., Bonal, M. T. L. y Díaz-Mata, M. A. (1994), Aprendizaje con redes neuronales artificiales. *Ensayos: Revista de la Facultad de Educación de Albacete*, 9, 169-180.

Rachudel, W. J. (1971). Multicollinearity once again. Harvard Institute of Economic Research. Cambridge.

Ramsey, J. B. y Schmidt, P. (1976). Some further results on the use of OLS and BLUS residuals in specification error tests. *Journal of the American statistical Association, 71*(354), 389-390.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics, 2*(1), 21-33.

Ridker, R., & Henning, J. (1967). The Determinants of Residential Property Values with Special Reference to Air Pollution. The Review of Economics and Statistics, 49(2), 246-257. doi:10.2307/1928231

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy, 82*(1), 34-55.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65(6): 386-408.

Rosenzweig, M. R., Leiman, L.A. y Breedlove, S. M. (2001). *Psicología biológica: Una Introducción a la neurociencia conductual, cognitiva y clínica*. Editorial Ariel, SA.

Rumelhart, D. E., et al. (1986). Learning representations by back-propagating errors. En Thad A. Polk y Colleen M. Seifert (Eds.), *Cognitive modeling* (pp. 213-220). Cambridge, Massachusetts.

Sanjuán, A. I., Hurlé, J. B., Pérez, L. y Royo, A. G. (2004). Análisis hedónico de los precios de la tierra en la provincia de Zaragoza. *Revista española de estudios agrosociales y pesqueros,* 202, 51-70.

Savin, N. y White, K. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. Econometrica, 45(8), 1989-1996. doi:10.2307/1914122.

Schibsted Classified Media (2015). Fotocasa.es: Alquiler de pisos, compra y venta. Recuperado de: https://www.fotocasa.es/es/

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

Shapiro, S. y Wilk, M. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591-611. doi:10.2307/2333709

Singh, I., Leitch, J. y Wilson, J., Gson User Guide, Google Inc., Recuperado de: https://github.com/google/gson

Slashdot Media (2016). Opencsv. Recuperado de: http://opencsv.sourceforge.net/

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics, 19*(2), 279-281.

Sturges, H. A. (1926), The choice of a class interval. *Journal of the American statistical Association, 21*(153), 65-66.

The Eclipse Foundation (2016). EclipseLink. Recuperado de: http://www.eclipse.org/eclipselink/

Trapletti, A. y Hornik, K. (2017). tseries: Time series analysis and computational finance. *R package version 0.10-42*. Recuperado de: https://cran.r-project.org/web/packages/tseries/tseries.pdf

Turing, A. M. (1936), Proceedings of the London Mathematical Society. *Proc. London Math. Soc.*, 42(1), 230-265.

Turing, A. (1950), Computing Machinery and Intelligence. *Mind*, 59(236), 433-460. Recuperado de: http://www.jstor.org/stable/2251299

Urbaniza Interactiva (2015). Portal inmobiliario Urbaniza.com Viviendas en venta y en alquiler. Pisos y casas a precio de mercado. Recuperado de: https://www.urbaniza.com/

Wallace, H. (1936). Farm Economists and Agricultural Planning. Journal of Farm Economics, 18(1), 1-11. Recuperado de: http://www.jstor.org/stable/1231308

Wang, Y., Potoglou, D., Orford, S., & Gong, Y. (2015). Bus stop, property price and land value tax: A multilevel hedonic analysis with quantile calibration. *Land Use Policy, 42,* 381-391.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48(4), 817-838. doi:10.2307/1912934

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural computation,* 1(4): 425-464.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*: Springer Science & Business Media. Recuperado de: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

Widrow, B. and Hoff M. E. (1960). Adaptive switching circuits. *Ire Wescon convention record,* 4(1): 96-104.

Yaencontre (2015). Yaencontre: pisos Madrid, pisos Barcelona, pisos alquiler, casas de compra, venta, alquiler y obra nueva. Recuperado de: https://www.yaencontre.com/