Universidad de
SanAndrés

Universidad de San Andrés

Departamento de Economía

Licenciatura en Economía

Real-Time Bidding: Predicción del comportamiento del precio de reserva en publicidad digital móvil

Autor: Sol Mizrahi Striebeck

Legajo: 25046

Mentor: Walter Sosa Escudero

Victoria, Buenos Aires, marzo 2018

ÍNDICE

Resumen

El presente trabajo tiene como objetivo predecir el precio de reserva del inventario publicitario digital móvil en el tiempo. La publicidad digital móvil es un medio de comunicación que expone a usuarios a anuncios de aplicaciones en dispositivos móviles. El mercado en el que se desarrolla se conoce como *Real-Time Bidding* (RTB) o *compra programática*, dado que implica la automatización de los procesos de compra y venta de inventario a través de subastas de segundo mejor precio en tiempo real. Se utiliza la base de datos del *Ad Exchange Inneractive* (plataforma tecnológica dónde ocurre el intercambio). El análisis se enfocará en el mercado de Estados Unidos para las fechas comprendidas entre el 25 de diciembre del 2015 y el 23 de marzo del 2017. El modelo final es un bucle que procesa las series de tiempo desagregadas por tipo y tamaño de activo por hora. Dados los patrones cíclicos de variación del precio de reserva, el sistema crea variables autorregresivas para cada una de las series y las evalúa a partir del procedimiento estadístico de *Análisis de Componentes Principales* (ACP) cada siete días. El bucle se complementa con un detector de *outliers* y valores ausentes. Luego, con el fin de suavizar los datos corruptos, el ruido blanco es intervenido a través de la función *spline*. Las predicciones de los valores futuros del precio de reserva se computan por hora. El modelo, al ser dinámico, logra adaptarse a las fluctuaciones del precio de reserva en el tiempo, mientras que toma en consideración la importancia de la granularidad de la información al estudiar las series individualmente. Se consigue estimar el precio de reserva con un error cuadrático medio promedio del 0.1999601, lo que representa un error medio absoluto relativo al precio de reserva de 3.58%.

1 Introducción

La publicidad digital móvil es un medio de comunicación que expone a usuarios a anuncios de aplicaciones en dispositivos móviles. Es una de las áreas que más ha crecido en la industria de la tecnología informática[1] en los últimos años. Según el reporte de ingresos publicitarios digitales[2] – publicado por el *Interactive Advertising Bureau* (IAB)[3] – los ingresos anuales del sector en 2016 llegaron a $72.5 mil millones de dólares en Estados Unidos, un incremento del 21.8% con respecto al 2015. Este crecimiento se vio impulsado, en su mayoría, por la publicidad móvil, que alcanzó por primera vez una participación de más del 50% de los ingresos publicitarios digitales.

Uno de los grandes sistemas involucrados en el desarrollo de la industria es del Real-Time Bidding (RTB), también conocido como *compra programática*, dado que involucra la automatización de los procesos de compra y venta del inventario publicitario digital. Las transacciones en RTB toman lugar a través de subastas de segundo mejor precio en tiempo real.

En busca de satisfacer un mercado en escala, RTB surge en el 2009 en lo que Yuan, Wang, Zhao (2013) definen como la intersección entre la liquidez de los datos y la del inventario publicitario. RTB es un concepto orientado al lado de la demanda, que utiliza algoritmos para automáticamente comprar y vender inventario en tiempo real. Estas transacciones ocurren en un lapso de unos 100 milisegundos desde que el Ad Exchange, (plataforma tecnológica que facilita la compra y venta de inventario digital), recibe la solicitud de publicación.

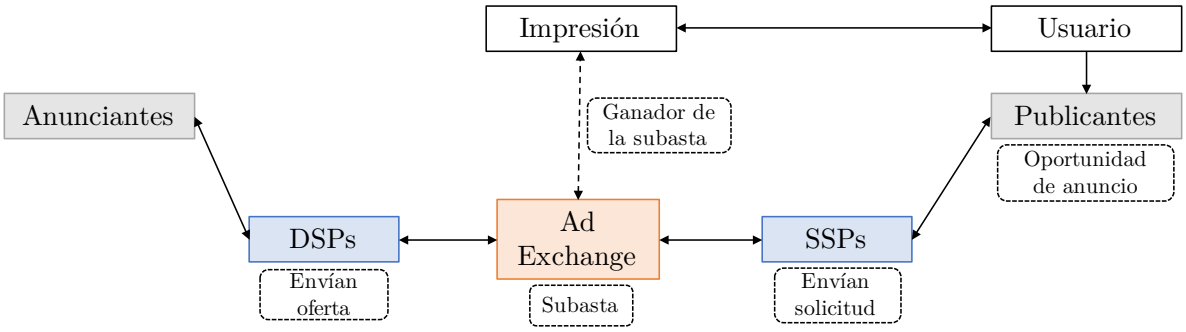
El mercado de la compra programática obedece en la siguiente estructura:

[1] También conocida como *IT* por sus siglas en inglés "information technology".

[2] En inglés, Internet Advertising Revenue Report, es una encuesta conducida de forma independiente por PricewaterCoopers (PwC) y publicado bianualmente al final del primer y tercer trimestre. Los resultados reportados son considerados los más precisos en la medición de la publicidad digital.

[3] El IAB, fundado en 1996, es una compañía que ayuda a las industrias de marketing y de medios a prosperar en la economía digital.

Diagrama 1: Estructura del mercado de RTB



Cuando un usuario interactúa con una aplicación móvil, una oportunidad publicitaria es creada. Inmediatamente, una plataforma de oferta (SSPs)[4] envía una 'solicitud de oferta' - en forma de subasta de segundo mejor precio - a las plataformas de demanda (DSPs)[5]. Esta solicitud es enviada a través de una plataforma de intercambio de anuncios, conocida como Ad Exchange. En respuesta, la DSP computa una 'oferta' para la subasta y la envía al Ad Exchange. Finalmente, de la subasta se obtiene al ganador, quien es notificado y su anuncio es expuesto al usuario que interactúa con la aplicación (Zhang, Yuan & Wang, 2014). La exposición de un usuario a un anuncio se denomina impresión.

Como la colocación de anuncios está determinada por subastas de segundo mejor precio, hay un foco muy grande en el comportamiento estratégico de los anunciantes y DSPs. Sin embargo, dada la complejidad y el dinamismo del proceso de la compra programática, es muy difícil aplicar estrategias predictivas efectivas que ayuden al proceso de optimización (Muthukrishnan, 2008).

La optimización de las estrategias de oferta es uno de los problemas más estudiados en la publicidad digital. El proceso de optimización en RTB consiste en tratar de maximizar el rendimiento de una campaña en base a un objetivo y sujeto a una restricción presupuestaria para un período determinado. Los objetivos, pueden ser especificados como la minimización

[4] Las plataformas de oferta o supply-side platforms (SSPs) son plataforma que controlan el inventario de las aplicaciones, es decir, la oferta publicitaria.

[5] Las plataformas de demanda o demand-side platforms (DSPs) son plataformas contratadas por anunciantes para administrar sus campañas publicitarias digitales, representando la demanda en este mercado.

de costo por clic[6] o costo por acción[7], así también como maximización del ratio de cliqueo[8] o ratio de acción[9] (Lee, Jalali & Dasdan, 2013).

Además, dado que el problema de optimización de las ofertas está sujeto a la restricción presupuestaria de la campaña, las estrategias resultan estar influidas directamente por el precio que efectivamente se paga si se gana la subasta y, por lo tanto, indirectamente por el precio al cual ofertan. Sin embargo, al estar inmersos en un sistema de subastas de segundo mejor precio, los apostadores ganadores pagan el monto ofertado por el segundo mejor postor. Esto implica que para que el anunciante conozca el vector de precios al que se enfrenta, necesita conocer la distribución de valoraciones privadas de sus competidores por una impresión en particular. Llegar a conocerla puede resultar en extremo difícil, no solo porque esta distribución puede ser muy compleja, sino porque puede ser que ni el competidor conozca con exactitud su valuación privada por cada impresión (Yuan, Wang et al., 2014).

No obstante, existe una noción del mínimo precio que se puede pagar en una subasta. El precio de reserva[10], o precio piso, se define como el mínimo precio por el que un publicador vende su espacio publicitario. En caso de que solo haya un postor, el precio que se paga es el precio piso. Al mismo tiempo, este mecanismo de exclusión de las subastas representa una limitación al poder de alcance de las campañas publicitarias, ya que, si la oferta es lo suficientemente baja, el anunciante ni llega a participar de la subasta.

También, puesto que los anunciantes buscan suavizar el gasto a lo largo del tiempo con el fin de lograr un impacto sustentable en el público, es fundamental para ellos definir el ritmo en el que ofertarán, mejor conocido como ritmo o frecuencia[11]. El ritmo o la frecuencia de una campaña es la porción de las solicitudes publicitarias disponibles por la que les gustaría ofertar para esta (Yuan, Wang & Zhao, 2013). Esto se da, a su vez, porque los anunciantes contemplan razones por las que deciden no ofertar de forma uniforme pero sí de forma dinámica.

[6] El costo por clic (CPC, por sus siglas en inglés "cost per click"), es el costo que paga el cliente por click atribuido.

[7] El costo por acción (CPA, por sus siglas en inglés "cost per action"), es el costo que paga el cliente por evento clave atribuido. Un evento es una acción que el usuario ejecuta a través de la aplicación móvil (una reserva, una compra, una suscripción, entre otros).

[8] El ratio de cliqueo (CTR, por sus siglas en inglés "click-through-ratio"), es el ratio de clics e impresiones.

[9] El ratio de acción (AR, por sus siglas en inglés "action ratio"), es el ratio de acciones y clics.

[10] Se representa al precio de reserva como el precio de reserva por cada mil impresiones.

[11] La frecuencia o ritmo es otro método utilizado en los procesos de optimización de oferta.

El objetivo final de ofertar dinámicamente es gastar más en aquellas horas que generan más clics o conversiones – o el evento con el que se optimiza la campaña. En consecuencia, podemos definir como '*tasa de éxito*' al ratio de impresiones y ofertas (Lee, Jalali & Dasdan, 2013). Este ratio, a su vez, es posible de mantener en el tiempo si sabemos el precio por el que es conveniente ofertar.

Simultáneamente, los anunciantes imponen precios máximos que pagan a las DSPs por clic, quienes tienen también sus propios objetivos de rentabilidad. Por ende, dentro de las variables internas de la campaña, los *bidders*[12] deben considerar sus márgenes. Así pues, como en toda industria, vemos metas contrapuestas que deben coexistir en este juego de estrategias protagonizado por los publicadores, los anunciantes y las plataformas de demanda.

Es así que la relevancia de conocer el precio de reserva se puede sintetizar con el siguiente argumento:

Dada la dificultad de conocer la valoración privada de los competidores, la definición de una frecuencia efectiva se ve obstaculizada. Además, al buscar la distribución de ofertas más costo-efectiva, se vuelve imprescindible obtener una noción sobre la estructura de precios a la que la demanda se enfrentará para lograr los objetivos de la campaña. Este vector de precios se contempla, en parte, a través del precio de reserva.

Es entonces que el objetivo de esta predicción es entender cómo el precio de reserva cambia en el tiempo. Para esto, se diseñó un bucle que procesa los datos del mercado y genera un modelo autorregresivo dinámico que computa predicciones del precio de reserva por hora. La aplicación del modelo busca evaluar las tendencias del mercado y asistir en el proceso de optimización de las campañas publicitarias digitales. En este caso, las bases de datos utilizadas pertenecen al Ad Exchange *Inneractive* para el mercado de los Estados Unidos. El inventario corresponde únicamente al de aplicaciones de dispositivos móviles.

El modelo resultante es un bucle que analiza series de tiempo. Está compuesto por 3 etapas. Primero, examina la presencia de intervalos sin observaciones, a los que definiremos como *valores ausentes*. Segundo, evalúa la existencia de *outliers* que estén corrompiendo el comportamiento cíclico de nuestra variable a predecir. Ambas fases se complementan con un tratamiento algorítmico (spline) que tiene como objetivo suavizar los datos corruptos. Por último, se crean variables autorregresivas y, mediante el método de *Análisis de Componentes Principales* (ACP), se seleccionan las variables a utilizar en las regresiones. El ACP es

[12] Plataforma licitadora de las DSPs con la que, a través algoritmos, definen por qué subastas de publicación ofertar y cuánto.

ejecutado cada 7 días para adaptarse a las fluctuaciones en el tiempo del precio de reserva. Las predicciones de los valores futuros del precio de reserva se computan por hora.

El trabajo se dividirá en 5 secciones: La segunda sección estará destinada a la exposición del estado del arte, en el que se discutirán las variables a considerar para el modelo predictivo y las dificultades metodológicas. La tercera sección, 'Datos y Metodología' especificará las bases de datos utilizadas y las metodologías y procesos aplicados en el cálculo de las predicciones. En la cuarta sección, se presentarán los resultados derivados del modelo y en la quinta, se evaluarán las conclusiones y posibles mejoras para futuras investigaciones.

2 Estado del arte

En esta sección se exponen los enfoques de diferentes publicaciones. Éstas asistirán en la determinación de las variables explicativas, los algoritmos y las herramientas que intervendrán en el diseño del modelo predictivo. De la misma forma, el estado del arte contribuirá a la discusión de las dificultades metodológicas y los consecuentes elementos a sopesar en el análisis.

2.1 Posibles dificultades metodológicas

El principal problema es la complejidad y dinamismo del ambiente en el que el Real-Time Bidding se desarrolla, haciendo que sea muy difícil aplicar estrategias de pronóstico efectivas. Por lo tanto, será relevante determinar cuáles son esos aspectos dinámicos a los que nos enfrentamos y, a su vez, comprender el universo en el que estaremos inmersos al analizar variables que dependen del correcto funcionamiento de la tecnología que las procesa.

En el intento de caracterizar el mercado de la compra programática, destacamos la existencia de numerosos estudios que han tratado calcular, entre otras cosas, el precio óptimo al cual ofertar. Adikari y Dutta (2015), por ejemplo, para lograr este objetivo proponen en su investigación una estrategia dinámica y autónoma para decidir sobre sus ofertas. Específicamente, se concentran en determinar el valor al que deben apostar, al mismo tiempo que intentan alcanzar los objetivos de la campaña. Las estrategias de oferta convencionales se basan en el número de apostadores y sus apuestas. En RTB, en cambio, el dinamismo existe en la cantidad de solicitudes de publicación recibidas de cada aplicación, los diferentes tipos de aplicaciones activas en un particular período de tiempo, el número de competidores

y los objetivos de la campaña de exposición[13] y/o gasto[14].

Como se mencionó anteriormente, el proceso de optimización en RTB consiste en tratar de maximizar el rendimiento de una campaña en base a un objetivo y sujeto a una restricción presupuestaria para un período determinado. Los objetivos, típicamente, pueden ser especificados como la minimización de costo por clic o costo por acción, así también como maximización del ratio de cliqueo o ratio de acción (Lee, Jalali & Dasdan, 2013). También las campañas pueden tener como objetivo escalar y ganar presencia en el mercado de la publicidad móvil. Este tipo de campañas conocidas como 'campañas de branding'.

Para evaluar la posibilidad de pronosticar la cantidad de solicitudes de oferta y el valor de las ofertas Adikari y Dutta utilizan un modelo autorregresivo integrado de media móvil o ARIMA[15]. El modelo es una versión generalizada del modelo autorregresivo de media móvil (ARMA), el cual es aplicable únicamente con series de tiempo. Debido al dinamismo del proceso de RTB, la exactitud de las predicciones es muy baja. Llegan a la conclusión de que, dado el inherente ecosistema en el que RTB se maneja y otras variables externas a este, pronosticar estas variables de acuerdo a valores históricos, no funcionará. Sin embargo, el desempeño del algoritmo puede ser ajustado reduciendo la duración del período entre ofertas.

Por su parte, Yuan, Wang y Zhao (2013) exponen que el valor por el cual ofertan depende de muchos factores, dada la impresión subastada. Puede estar influido tanto por los KPIs[16] que se predice que se obtendrán, la restricción presupuestaria, la probabilidad de ganar la subasta y características y costos particulares de la impresión.

Además, si bien los anunciantes buscan suavizar el gasto a lo largo del tiempo con el fin de lograr un impacto sustentable en el público, los autores consideran que hay razones por las

[13] Los objetivos de exposición determinan cómo se exhibirán los anuncios durante la campaña. Algunos de los aspectos a definir son la segmentación de base de usuarios de acuerdo a su comportamiento, el mensaje que se envía al usuario de acuerdo al segmento al que pertenece, frecuencia de impresión, entre otros.

[14] Los objetivos de gasto determinan la restricción presupuestaria de la campaña.

[15] Acrónimo del inglés (Autoregressive Integrated Moving Average). ARIMA es un modelo estadístico que utiliza las observaciones de series de tiempo para predecir futuras tendencias. Es un tipo de regresión que busca predecir futuras fluctuaciones. Los rezagos de las series de tiempo diferenciadas se denominan 'autorregresivos' y rezagos dentro de valores estimados se denominan 'media móvil'. Este tipo de modelo puede considerar tendencias, estacionalidad, ciclos, errores y aspectos no estacionarios de los datos cuando genera predicciones (disponible en http://www.investopedia.com).

[16] Acrónimo de Key Perfomance Indicator. Un KPI es un indicador clave de rendimiento, es decir, un valor que es posible de medir y que define qué tan efectivamente un negocio cumple con sus objetivos claves.

que se decide no ofertar de forma uniforme pero sí de forma dinámica. Estos fundamentos definen la regularidad con la que las DSPs ofertarán, mejor conocido como ritmo o frecuencia. La frecuencia de una campaña es la porción de las solicitudes publicitarias disponibles por la que nos gustaría ofertar para esta.

En definitiva, remarcan que ciertos problemas de tráfico pueden derivar en que el ofertar uniformemente a lo largo del día sea una estrategia ineficiente. Si la mayoría de las impresiones de calidad aparecen al principio del día, la configuración no podrá capturar todas estas; en tanto, si no hay el suficiente tráfico al final del día, la configuración tampoco permitirá gastar todo el presupuesto, llevando a una situación de sub-provisión. Por lo tanto, es lógico pensar en que se elegirá ofertar de forma dinámica para gastar más en aquellas horas que generan más clics o conversiones.

Al mismo tiempo, si bien la teoría de subasta óptima para subastas de segundo mejor precio nos indica que sin importar la existencia de un precio de reserva los bidders están incentivados a ofertar sus valoraciones privadas, hay que notar que esta estrategia dominante no se sostiene en las publicidades actuales. Yuan, Wang et al. (2014) señalan que una de las principales razones es que, generalmente, se utilizan puntajes de calidad para clasificar las impresiones con determinados publicadores. Sin estos rankings, la estrategia dominante de ofertar sus valoraciones privadas forma parte del equilibrio de Nash del sistema significando que, conforme pasa el tiempo, los anunciantes no tienen incentivos de cambiar sus ofertas, ceteris paribus.

Por otro lado, Xiao, Yang y Li (2009) explican que los publicadores muy a menudo utilizan precios de reserva para incentivar a los anunciantes al ofertar. Esta es una estrategia atractiva dado que es muy fácil ajustar los precios y, por lo tanto, controlar los ingresos. En definitiva, si el precio es muy alto, los anunciantes no podrán ni participar de la subasta y los ingresos se verán afectados al dejar inventario desperdiciado. En contraparte, si el precio es muy bajo, se perderá la oportunidad de aumentar los ingresos, ya que los anunciantes estarían dispuestos a pagar más. Es así que los autores establecen que ciertas variables, como el número de anunciantes ofertando, generan un impacto en el precio de reserva óptimo.

Simultáneamente, los anunciantes imponen precios máximos que pagan a las DSPs por clic y éstas, a su vez, tienen metas de rentabilidad. Por ende, dentro de las variables internas de la campaña, los bidders deben considerar sus márgenes. Vemos entonces, metas contrapuestas que deben coexistir en este juego de estrategias protagonizado por los publicadores, los anunciantes y las DSPs.

Así, dada la dificultad de conocer la valoración privada de los competidores, la definición de una frecuencia efectiva se ve obstaculizada. Por otro lado, al buscar la distribución de la

estrategia de oferta más costo-efectiva se vuelve imprescindible obtener una noción sobre qué estructura de precios se enfrentará para que la demanda logre su renta deseada, al mismo tiempo que entregue el volumen y la calidad requeridas por su cliente. Este vector de precios en el mundo dinámico de la compra programática se contempla, en parte, a través del precio de reserva ya que representa las tendencias en el mercado, volviendo relevante el conocer sus fluctuaciones en el tiempo.

Por lo tanto, luego de esta primera instancia del estudio, es posible destacar 2 características del RTB que se deberán afrontar: un alto número de variables que intervienen en su variación y su dinamismo.

2.2 Granularidad: problemas de dimensionalidad

Dada la cantidad de posibles variables que pueden influir en la fluctuación del piso de reserva, el modelo puede potencialmente sufrir de problemas de dimensionalidad. Para evitar este tipo de problemas, se implementará el análisis de componentes principales para la selección de variables.

Siguiendo a James et al. (2013), el análisis de componentes principales consiste en construir M componentes principales, $Z_1, \dots, Z_M$, a través de la mejor combinación lineal (la de máxima varianza) de las p variables del modelo ($M < p$). Es un método no supervisado ya que solo involucra un conjunto de características $X_1, \dots, X_p$ (variables explicativas), no asociadas a la respuesta Y (variable explicada). A su vez, luego es posible usar estos componentes como predictores en un modelo de regresión lineal que ajuste por mínimos cuadrados.

La idea viene por parte de que, normalmente, un número pequeño de componentes es suficiente para explicar la mayor parte de la variabilidad de los datos, así también como su relación con la "respuesta". Es decir, vamos a asumir que las direcciones en las que variables explicativas $X_1, \dots, X_p$ muestran mayor variación, son direcciones asociadas con la variable explicada Y.

Además, en el proceso se estandarizan los predictores. La estandarización asegura que todas las variables estén en la misma escala y es un proceso similar al promedio, con la diferencia en que no restringe la existencia de ponderaciones negativas, por lo que es una suerte de índice. Con lo cual, se asume que si se calculan $M \ll p$ componentes (donde M se elige utilizando *cross-validation*), se evitarán problemas de dimensionalidad como el sobreajuste. A diferencia de MCO, es un método no supervisado porque no hay Y que guíe el proceso, sino que Y es la que se construye con cada componente. Esto significa que el primer componente principal es la mejor forma de representar Y usando una sola variable. En tanto,

el segundo componente principal es la mejor combinación lineal ortogonal a la mejor inicial, y así progresivamente.

El primer componente principal del conjunto de variables $X_1, \dots, X_p$ es la combinación lineal normalizada de las variables que es la de máxima varianza:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Los elementos $\phi_{11}, \dots, \phi_{p1}$ son los coeficientes del primer componente principal y juntos conforman el primer vector de coeficientes del análisis de componentes principales $\phi_1 = (\phi_{11}\ \phi_{21} \dots \phi_{p1})^T$. Por normalizado, implicamos que $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Al imponer esta restricción, evitamos que estos elementos sean arbitrariamente grandes en valor absoluto, lo que podría resultar en una varianza arbitrariamente grande.

Para computar el primer componente principal, dado que estamos únicamente interesados en la varianza, asumimos que cada una de las variables X ha sido centrada para tener promedio cero, i.e. la columna de medias de X es cero. Luego, buscamos la combinación lineal de la muestra de valores:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

que tenga la mayor varianza sujeto a la restricción $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. En otras palabras, el vector de coeficientes del primer componente principal resuelve el problema de optimización:

$$\max_{\phi_{11},\dots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \ sujeto\ a \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Donde la función objetivo puede ser reescrita como $\frac{1}{n}\sum_{i=1}^{n}(z_{i1})^2$. Como $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, el promedio de $z_{11}, \dots, z_{n1}$ va a ser cero también. Además, la función objetivo que se maximiza es solo una muestra de la varianza de los n valores de z_{i1}. Esto significa que la función objetivo maximiza la variabilidad de una representación de X. Nos referimos a $z_{11}, \dots, z_{n1}$ como puntajes del primer componente principal.

Este primer componente principal tiene, como propiedad, que se encuentra en la línea en el espacio de dimensión p que está más cerca de las n observaciones (en términos de distancia Euclídea cuadrada promedio). Esto sucede porque Z_1 busca una dimensión que resuma lo mejor posible todos los datos.

Por otro lado, el vector de coeficientes ϕ_1 con los elementos $\phi_{11}, \dots, \phi_{p1}$ define una dirección en el espacio en el que haya mayor varianza. Si proyectamos las n observaciones $x_1, \dots, x_p$ en esa dirección, estas proyecciones son los puntajes de los componentes principales $z_{11}, \dots, z_{n1}$.

Luego de determinar el primer componente principal Z_1, se debe hallar el segundo. Éste es la combinación lineal de $X_1, \dots, X_p$ con máxima varianza dentro de todas las posibles combinaciones lineales que *no estén correlacionadas* con Z_1. Los puntajes del segundo componente principal $z_{12}, z_{22}, \dots, z_{n2}$ toman la siguiente forma

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}$$

donde ϕ_2 es el vector de coeficientes del segundo componente. El vector de coeficientes del segundo componente principal resuelve el problema de optimización:

$$\max_{\phi_{12},\dots,\phi_{p2}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j2}x_{ij}\right)^2 \right\} \; sujeto\ a\ \ (1)\sum_{j=1}^{p}\phi_{j2}^2 = 1 \ \ y\ a\ \ (2)\, Cov(Z_1, Z_2) = 0$$

El restringir a Z_2 a no estar correlacionada con Z_1 es el equivalente a restringir la dirección ϕ_2 a ser ortogonal a la dirección de ϕ_1[17].

Análogamente, se calculan los $M - 2$ componentes principales restantes. El resultado final es la construcción de un conjunto de combinaciones lineales ortogonales que pierden información sucesivamente.

2.3 Dinamismo: tratamiento de ruido blanco en series de tiempo

El segundo aspecto de las series de tiempo a analizar es el dinamismo del ambiente en el que interactúan. Para las variables conocidas y cuantificables, hemos determinado contemplarlas a través de componentes principales. Sin embargo, existe un número de variables desconocidas que intervienen en la fluctuación del precio de reserva perturbando su comportamiento.

Al tratar con un modelo autorregresivo, existe la necesidad de lidiar con estas variables desconocidas, al menos de forma indirecta. Determinamos que es menester incluir un sistema

[17] Las direcciones de los componentes principales $\phi_1, \phi_2, \dots, \phi_p$ son la secuencia ordenada de los autovectores de la matriz X^TX y las varianzas de los componentes son sus autovalores.

capaz de detectar como también tratar la presencia de *outliers* para evitar que datos corruptos alteren la precisión predictiva del modelo. En esta oportunidad, el tratamiento se aplicará a través del método de spline.

Siguiendo a Gobron N. et al. (2011) spline es un método que proporciona una forma atractiva de suavizar los datos corruptos observados en N puntos arbitrariamente localizados en un intervalo finito de tiempo. Este método no asume causas subyacentes de las variaciones ni en la estructura matemática de la serie. Spline construye una curva continua de segmentos de polinomios cúbicos agrupados en puntos de manera que la primera y la segunda derivadas de la curva resultante sea completamente continua. Este método es aplicable a un amplio rango de bases de datos porque es flexible (hace pocas suposiciones) y es ajustable a través de un solo parámetro de ajuste que controla la "rigidez" o la "flexibilidad" de la curva del spline.

James et al. (2013) indica que un abordaje para spline es el de encontrar la función g que minimice

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

$\lambda > 0$ es el parámetro de ajuste y se elige por cross-validation con el fin de minimizar la suma de cuadrados de los residuos.

La función g se conoce como *spline suavizador*. El primer término $\sum_{i=1}^{n}(y_i - g(x_i))^2$ es la *función de pérdida* que lleva a g a ajustar bien los datos.

En contraparte, el término $\int g''(t)^2 dt$ es el *término de penalidad* que penaliza la variabilidad de g – dada por la segunda derivada de g que es la que indica la cantidad en la que la pendiente de la curva está cambiando en t (primera derivada de g). En definitiva, el término $\int g''(t)^2 dt$ es la medida del cambio total en la función $g'(t)$ en todo su rango. Si g es muy suave, entonces $g'(t)$ será cercano a una constante y $\int g''(t)^2 dt$ tendrá un valor cercano a cero. Contrariamente, si g es variable entonces $g'(t)$ variará de forma significativa y $\int g''(t)^2 dt$ tomará un valor alto. Por lo tanto, $\lambda \int g''(t)^2 dt$ incita a g a ser suave. Cuanto más alto sea el valor de λ, más suave será g.

Para valores pequeños de λ, el spline permanece cercano a los puntos de los datos y, en el caso límite, cuando $\lambda = 0$ el término de penalidad no tiene efecto y la función simplemente interpola los datos. Por el contrario, altos valores de λ incrementan la rigidez de la curva y, en el caso límite, cuando $\lambda \to \infty$, g es perfectamente suave y el spline se vuelve una regresión de mínimos cuadrados.

Al ser spline un método simple, robusto y computacionalmente no costoso, resulta apropiado para procesar bases de datos grandes y se utilizará como tratamiento para suavizar los datos ante presencia de outliers y valores ausentes (Gobron N. et al., 2011).

2.4 Aplicación al caso de estudio

De aquí concluimos que, con el fin de alcanzar el objetivo propuesto, será prioritario utilizar la mayor granularidad posible para nuestro modelo. Si bien ciertas variables no serán posibles de contemplar dentro del cálculo, es importante adquirir conciencia de su existencia y posible impacto en el análisis; e.g. los diversos componentes que influyen en la decisión de los anunciantes van a ofertar, afectando indirectamente al precio de reserva en el largo plazo. Por otro lado, existen una serie de factores que se deben considerar al evaluar la fluctuación del precio de reserva en el tiempo como el comportamiento diario y la estacionalidad.

Por ende, dada la cantidad de posibles variables que se podrían tomar en cuenta para la regresión, el método de componentes principales parece ser adecuado para su selección. Con este método evitaremos problemas de dimensionalidad y podremos determinar las variables que expliquen mejor y en mayor proporción a la variabilidad del precio de reserva. Así también como definir el orden de relevancia con el que la explican.

Por último, se ha evaluado la posible corrupción de los datos derivada tanto del dinamismo de RTB como de posibles incidentes tecnológicos. Al tratarse de un modelo autorregresivo, el error en las predicciones no solo se ve influido por la presencia de ruido blanco en la observación analizada, sino también por la utilización de datos corruptos para la estimación de futuras observaciones. Es entonces que, se aplicará spline para suavizar los datos ruidosos o faltantes a fin de estimar valores que los reemplacen suavizando la serie de tiempo.

3 Datos y metodología

En esta sección se especifican las bases de datos utilizadas y las metodologías y procesos aplicados en el cálculo de las predicciones.

3.1 Datos

Se hará uso de las bases de datos pertenecientes a la plataforma analítica en tiempo real para la publicidad digital desarrollada por Metamarkets Group Inc. Esta plataforma otorga la posibilidad de visualizar tendencias y oportunidades en el mercado programático.

Específicamente, dado que el vector de precios al que se enfrenta un bidder varía para cada

país, se empleará la base de datos del Ad Exchange *Inneractive* y el análisis se enfocará en el mercado de Estados Unidos. Las series de tiempo seleccionadas se fraccionan por hora para las fechas comprendidas entre el 25 de diciembre del 2015 hasta el 23 de marzo del 2017. Además, están desagregadas por tipo y tamaño de creativo – 2 variables características del espacio publicitario. El inventario que utilizaremos se encuentra disponible únicamente en aplicaciones de dispositivos móviles.

3.2 Metodología

El objetivo del modelo predictivo es evaluar tendencias de mercado y entender cómo el precio de reserva cambia en el tiempo.

En general, los bidders toman una selección amplia de variables para realizar un análisis granular de la información. Al mismo tiempo, buscan adaptar y perfeccionar sus modelos haciendo uso de las observaciones históricas de la campaña para generar nuevas y más precisas predicciones sobre su valuación por cada impresión. Por otro lado, al analizar el comportamiento del precio de reserva, se observan patrones cíclicos en su fluctuación en el tiempo. Como este movimiento se percibe diariamente, la hora del día se define como un factor fundamental en este estudio.

Por lo tanto, dada la conducta del mercado y que tratamos con predicciones de series de tiempo, se concluyó construir un modelo autorregresivo. A partir de este concepto, determinamos a las observaciones históricas del precio de reserva o "rezagos" como las primeras variables explicativas del modelo.

El utilizar datos históricos, asimismo, permitirá contemplar la condición de estacionalidad[18] de la que el mundo del comercio digital se comprende. Tales tendencias suelen contribuir al aumento del precio piso dado que los anunciantes están dispuestos a pagar más por espacios publicitarios con el fin de ganas competitividad e incrementar sus ventas.

Al mismo tiempo, como Xiao, Yang y Li (2009) explican, el número de anunciantes ofertando puede generar un impacto en el precio de reserva óptimo. De modo que el número de ofertas que recibe el Ad Exchange también se considerará.

Además, las regresiones deberán evaluarse a nivel global y por tamaño de creativo, dada la discrepancia en niveles de precios y fluctuaciones entre tamaños y/o tipos de activos *(ver*

[18] Meses en el año o días festivos, por ejemplo.

Anexo I). Esto permitirá a su vez estudiar el rol de la granularidad de los datos en las predicciones.

Sin embargo, dada la cantidad de variables que se podrían tomar en cuenta para la regresión, en especial al examinar las posibles variables autorregresivas, se determinó aplicar el método de componentes principales para su selección. De este modo, evitaremos problemas de dimensionalidad como el sobreajuste y podremos determinar las variables que expliquen mejor y en mayor proporción a la variabilidad del precio de reserva. Es decir, que también podremos definir el orden de relevancia con el que explican esta variabilidad. El número de componentes principales se selecciona por cross-validation.

Por otro lado, al evaluar la naturaleza de RTB, establecemos la necesidad de recalcular el análisis de componentes principales en distintos lapsos en el tiempo para flexibilizar el modelo y adaptarlo al dinamismo que lo caracteriza. Al mismo tiempo, parece prudente contemplar la idea de que datos demasiado antiguos no revelan información significativa para a predicciones ya que el mercado evoluciona y varía en el tiempo. Esto significa que también se deberá definir la ventana de datos históricos sobre la que se calcularán los componentes principales y el modelo predictivo.

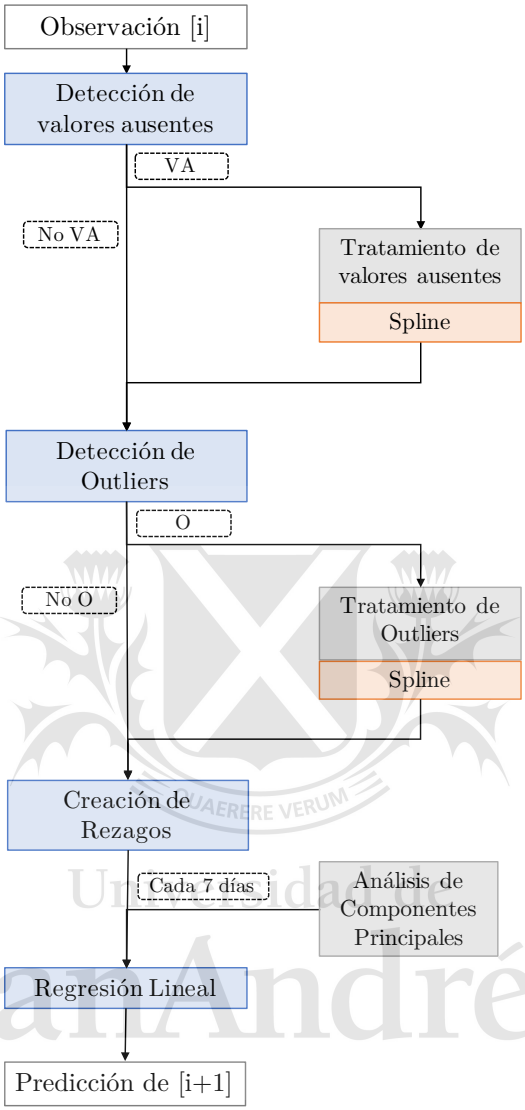
Sumando a los elementos mencionados anteriormente, y tratándose de un modelo autorregresivo, es indispensable comprender el efecto que los outliers pueden tener en las predicciones. Al basar las predicciones en datos históricos, una observación que resulte estar fuera del rango de sus valores esperados, no solo puede afectar el error de predicción para dicha observación sino, también, el de las siguientes. Por lo tanto, se sopesará la relevancia de este posible inconveniente implementándose un sistema de detección y tratamiento de outliers.

De la misma forma, consideraremos dentro de nuestro mundo de posibilidades un posible caso en que un servidor falle – ya sea el del Exchange o el de una DSP – o que haya otro tipo de desperfecto técnico que no permita leer los datos creando un intervalo en el tiempo sin observaciones. Se ejecutará entonces un sistema de detección y tratamiento de estos valores ausentes en la serie de tiempo.

Ambos tratamientos se ejecutarán a través de la función spline. Al detectar la ausencia de un valor o un outlier, spline computará un valor que reemplace la observación corrupta suavizando los datos con los que se calcularán los componentes principales y las consecuentes predicciones.

El bucle final obedece la siguiente estructura:

Diagrama 2: Estructura del bucle



Entre observaciones, primero se verifica que no haya ausencia de observaciones en la serie. En caso de haberlo, se introduce una fila con *NAs* (valores nulos). Luego, spline computa un valor para reemplazar el valor nulo para las series de "solicitudes de oferta" y del "precio de reserva promedio".

Posteriormente, se analiza la presencia de outliers para las series del precio de reserva promedio y las solicitudes de oferta. Si el valor de la observación resulta ser 1.5 veces más grande *(0.5 veces más chica)* que las observaciones históricas[19] y la variación porcentual de

[19] Las observaciones correspondientes al mismo horario de la observación evaluada de los 7 días anteriores.

la observación con respecto a la de la hora anterior resulta ser mayor *(menor)* que la variación porcentual histórica[20], entonces la observación se clasifica como outlier y se calcula un valor alternativo a través de spline.

Una vez procesados los datos, se crean los rezagos correspondientes para esta observación y, por último, se corre la regresión para predecir i+1 (la hora siguiente) con los componentes que derivan del análisis de componentes principales. ACP se estima cada siete días y la cantidad de componentes seleccionados se utiliza para las estimaciones de esa semana.

4 Resultados

En esta sección se presentarán los resultados derivados del análisis de los datos y las estimaciones del modelo.

4.1 Valores ausentes y outliers

Comenzando con el procesamiento de los datos, observamos los efectos de aplicar un sistema de detección de outliers y valores ausentes. En conjunto, advertimos la efectividad de spline para suavizar los datos corruptos en las series de tiempo.

4.1.1 Valores ausentes

En total, se hallaron 5 valores ausentes en dos intervalos de tiempo en las series, es decir, todas las series carecían de datos para estos lapsos. El primer intervalo se encontró en la fecha del 6 de enero del 2016, en el horario desde las 11:00:00 hasta las 15:00:00. El segundo intervalo se halló en la fecha del 15 de noviembre del 2016, en el horario desde la 01:00:00 hasta las 04:00:00.

4.1.2 Outliers

La tabla a continuación presenta la cantidad de outliers encontrados por tamaño/tipo de activo en las variables del precio de reserva promedio y solicitudes de oferta.

[20] Las variaciones porcentuales entre observaciones correspondientes al mismo horario de los 7 días anteriores.

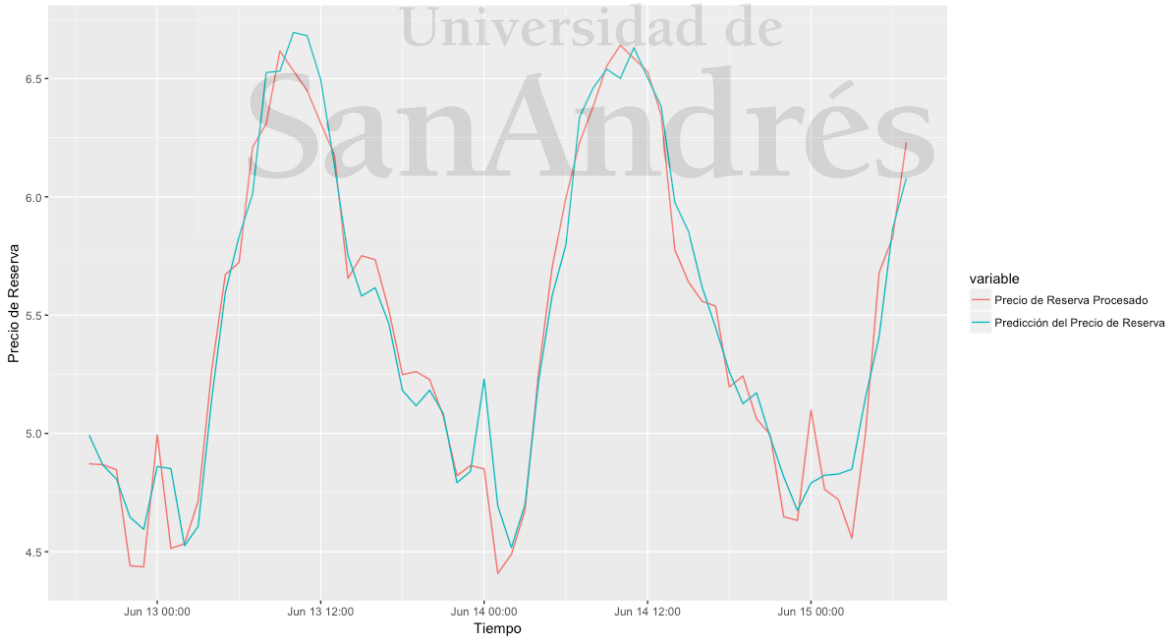
Tabla 1: Outliers hallados por serie de tiempo

Serie	*Nº de Outliers (Precio de reserva)*	*Nº de Outliers (Solicitudes de oferta)*
Global	18	37
768x1024	63	53
1024x768	14	51
320x480	52	39
480x320	25	87
300x250	55	34
728x90	2	29
320x50	20	25
Video	0	86

Gracias a este proceso, la base de datos alcanza un patrón cíclico sin presencia de ruido blanco que permite crear posteriormente variables autorregresivas "suavizadas" *(ver Anexo II).*

Consecuentemente, se logra calcular predicciones acordes a una base sin valores corruptos. El siguiente gráfico es una representación de los datos procesados y su correspondiente predicción a través del análisis de componentes principales ejecutado cada 7 días:

Gráfico 1: Datos procesados vs. predicción – Serie global



4.2 Error cuadrático medio y error medio absoluto

En segundo lugar, se analiza la efectividad del modelo para predecir el precio de reserva en el tiempo.

En definitiva, a través del sistema de detección de intervalos de valores ausentes y outliers, se ha logrado estimar el precio de reserva con un error cuadrático medio de 0.1130771[21] para el caso de la serie global[22]. El suavizar las observaciones implica una gran reducción del error cuadrático medio. Específicamente, al estudiar esta serie sin evaluar estas condiciones ni procesar la información, se halla que la serie global deriva en un error cuadrático medio de 2.388298e+32.

También, se estimaron las predicciones con distintas series – dada la discrepancia en niveles de precios y fluctuaciones entre tamaños y/o tipos de activos *(ver Anexo I)* - con el fin de estudiar el rol de la granularidad de los datos en las predicciones. Además, con la intención de evaluar la relevancia de los datos históricos en el tiempo, se corrió el modelo utilizando distintas ventanas para el cálculo de los componentes principales. Este procedimiento se llevó a cabo para buscar establecer el tamaño de muestra sobre el que el análisis de componentes principales se debe ejecutar.

A continuación, se presenta el error cuadrático medio para cada modelo estimado. Se desglosa la información por tamaño de imagen o tipo de activo para distintas ventanas de tiempo. En negrita se remarca la ventana de observaciones históricas que minimiza el error cuadrático medio para cada serie.

[21] El cálculo del error cuadrático medio se estimó utilizando las predicciones y los datos procesada neta de outliers y ausencia de observaciones, es decir, neta de valores corrupta. Al calcular el error cuadrático medio con la base neta de ruido blanco se ignora el efecto que los outliers y ausencia de observaciones pueden tener sobre la estimación del modelo.

[22] Promedio de todos los precios de reserva de la base de datos.

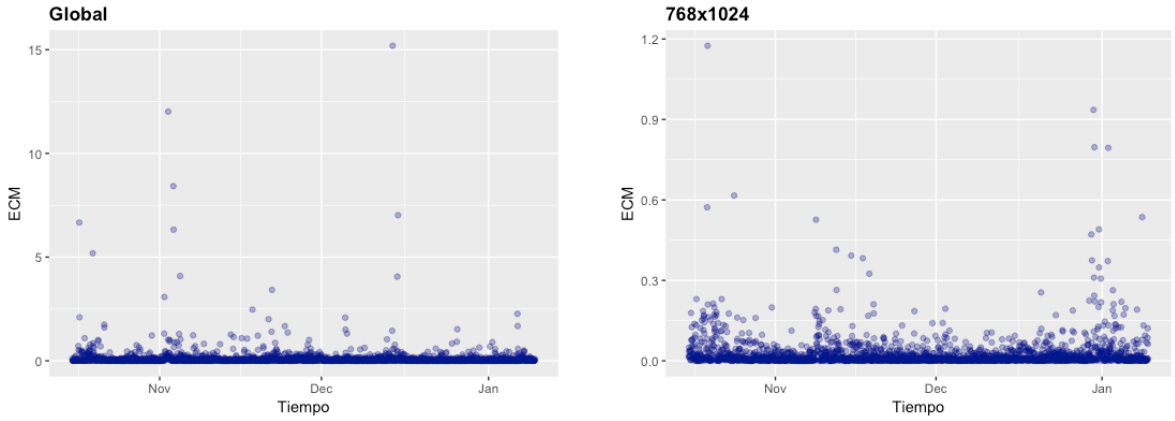
Tabla 2: Error cuadrático medio por serie de tiempo

Serie	*30 días*	*60 días*	*90 días*	*120 días*	*150 días*	*180 días*
Global	0.122270	**0.113077**	0.117328	0.115917	0.116386	0.116057
768x1024	**102.0510**	131.7542	132.2377	132.0320	132.3461	132.3507
1024x768	0.116473	**0.106785**	0.112341	0.110221	0.108941	0.107032
320x480	3.616887	3.612260	**3.322575**	3.325833	3.337188	3.323566
480x320	0.401204	0.330495	0.320776	0.320940	**0.315312**	0.318106
300x250	**0.772269**	0.981671	0.979168	0.979965	1.007711	1.006419
728x90	0.003899	0.003722	0.003635	0.003583	0.003492	**0.003466**
320x50	0.036483	0.033496	0.032734	**0.031992**	0.032440	0.032236
Video	0.065603	0.060926	0.059542	0.058962	0.058407	**0.056820**

En una primeria instancia, se observa que la reducción del error cuadrático medio con respecto al estudio de la serie global se da en la mitad de los casos. Sin embargo, el mayor beneficio de explorar este grado de granularidad radica principalmente en que es posible analizar la variabilidad del precio de cada serie por separado. Si bien se halla un comportamiento similar entre las series de carácter cíclico, en cada una nos enfrentamos a distintos niveles precios y, por lo tanto, a variabilidades de precios distintas. Esto significa que la divergencia entre errores cuadráticos medios es reflejo de analizar cada rango por separado. En consecuencia, las predicciones resultantes son más efectivas en la práctica ya que se podrán utilizar para ofertar a un precio más adecuado al tipo y tamaño de espacio publicitario.

A continuación, se graficó el menor error cuadrático medio alcanzado en cada serie de tiempo:

Gráfico 2: Menor error cuadrático medio alcanzado



Al mismo tiempo, se observa un mayor error cuadrático medio en las series de 320x480 y 768x1024, reflejando las dificultades del modelo al operar con series con una mayor cantidad de outliers. Esto sucede porque, al aumentar la cantidad de outliers, las series pierden su carácter de periodicidad y esta característica es la que, precisamente, vuelve al modelo autorregresivo efectivo para la predicción de valores futuros.

Examinando la base de datos en mayor profundidad, se halla que en dichas series ocurren casos aleatorios de extrema divergencia entre el dato procesado y el dato predicho. Es decir, spline falla en corregir los valores detectados como outliers en estos casos, sesgando el valor del error cuadrático medio *(ver Anexo III y IV).* Sin estos pocos errores en los datos procesados, el error cuadrático medio se asimilaría al estimado en el resto de las series.

De igual manera, es relevante comprender cuál es el porcentaje del error medio absoluto (EMA) del modelo. La *Tabla 3* resume el error medio absoluto correspondiente a la ventana que minimiza el error cuadrático medio por serie.

Tabla 3: Resumen del error cuadrático medio y error medio absoluto por serie de creativo

Serie	*Precio Medio*	*Ventana*	*ECM*	*EMA*	*% EMA*
Global	6.007318	60 días	0.113077	0.226438	3.769362
768x1024	5.580572	30 días	102.0510	1.113181	19.94743
1024x768	4.541849	60 días	0.106785	0.194839	4.289865
320x480	5.510749	90 días	3.322575	0.554068	10.05432
480x320	7.176750	150 días	0.315312	0.282555	3.937091
300x250	2.051166	30 días	0.772269	0.180951	8.821837
728x90	1.746874	180 días	0.003466	0.036192	2.071797
320x50	2.043873	120 días	0.031992	0.067205	3.288135
Video	8.304210	180 días	0.056820	0.153181	1.844612

Observamos para la serie global un EMA de $0.23 que representa un 3.77% del precio medio. La serie 768x1024 alcanza el mayor error medio absoluto relativo al precio con un 19.95% y la serie de video el menor con un 1.85%. La diferencia en puntos porcentuales del error medio absoluto relativo al precio entre cada serie, en parte, es reflejo de la divergencia en la cantidad de outliers detectados. Las series con mayor porcentaje de error medio absoluto relativo al precio (768x1024, 320x480 y 300x250), a su vez, son las más corruptas. Específicamente, la cantidad de outliers hallados en el precio de reserva de estas series es del doble que del resto.

4.3 Determinación de la ventana de tiempo para el análisis de componentes principales

La ventana de observaciones óptima para computar el análisis de componentes principales cada 7 días es el segundo elemento que se buscó determinar. Apreciando la *Tabla 2* nuevamente, se advierte la inexistencia de una ventana de tiempo que minimice el error cuadrático medio para todas las series. Se podría concluir que este debería ser otro aspecto dinámico a tomar en cuenta en el desarrollo de un bucle más sofisticado.

4.4 Estructura de los componentes principales

Por otro lado, evaluamos la estructura de los componentes principales del modelo. Se observa que, en los últimos análisis calculados por el bucle, los primeros 10 componentes generalmente logran explicar aproximadamente entre un 85-90% de la variabilidad del precio de reserva *(ver Anexo V).* Sin embargo, dado el dinamismo del modelo, esta característica varía en el tiempo.

Además, nos percatamos de que[23] los primeros 1 a 3 componentes del modelo están integrados por ponderaciones negativas de los 3 rezagos anteriores y posteriores a la hora [i+1]-12 y [i+1]-36[24], y ponderaciones positivas de 3 rezagos anteriores y posteriores a la hora [i+1]-24 y [i+1]-48[25]. Esta fluctuación en el tiempo es coherente con el aspecto cíclico de la serie de tiempo *(ver Anexo VI).*

Por último, advertimos que los componentes siguientes están ponderados cada vez en menor medida por los valores de los rezagos anteriores y posteriores cercanos a las horas [i+1]-12 y [i+1]-36. Al mismo tiempo, cuanto más alto sea el número del componente, estará ponderado en mayor medida por los rezagos anteriores y posteriores cercanos a las horas [i+1]-24 y [i+1]-48.

5 Conclusiones

El dinamismo del Real-Time Bidding genera varias incógnitas en la industria sobre las mejores prácticas en los procesos de optimización de las campañas publicitarias. Las estrategias para ofertar forman parte de estos procesos. Dada la dificultad de saber la cantidad de competidores y sus valoraciones privadas, se desconoce el vector de precios al

[23] de acuerdo al último análisis de componentes evaluado en la serie global

[24] Valores de 12 horas anteriores al horario a predecir de días anteriores.

[25] Valores en el mismo horario a predecir de días anteriores.

que nos enfrentamos en cada subasta y, por lo tanto, el valor al que es conveniente ofertar. Sin embargo, a través del precio de reserva es posible tener una comprensión más efectiva de las fluctuaciones del mercado.

El modelo construido en forma de bucle ejecuta predicciones sobre el precio de reserva, empleando como método de selección de variables el análisis de componentes principales. Al recalcular periódicamente el ACP y automatizar la selección de componentes a utilizar en las regresiones, se logra flexibilizar el modelo adaptándolo a la variabilidad estacional del precio de reserva.

Además, observamos que, de acuerdo al último análisis de componentes evaluado en la serie global, los primeros 1 a 3 componentes del modelo están integrados por ponderaciones negativas de los 3 rezagos anteriores y posteriores a las horas [i+1]-12 y [i+1]-36. En contraparte, también están conformados por ponderaciones positivas de los 3 rezagos anteriores y posteriores a las horas [i+1]-24 y [i+1]-48. Esta composición es coherente con el aspecto cíclico de las series estudiadas e implica que nuestro modelo logra captar esta característica a través de las variables autorregresivas.

Por otro lado, en las series evaluadas, también se halla ruido blanco. La corrupción en las series de tiempo representa un gran desafío durante el desarrollo de un análisis autorregresivo ya que, aleatoriamente, perturba las estimaciones disminuyendo la eficacia de las predicciones presentes y futuras. Es por eso que, al tratar con series de tiempo, se vuelve indispensable contar con un sistema de detección y tratamiento de observaciones ruidosas.

La cantidad de outliers detectados impacta al error medio absoluto relativo al precio de cada serie. Las series con mayor porcentaje de error medio absoluto relativo al precio (768x1024, 320x480 y 300x250), a su vez, son las más corruptas. Específicamente, la cantidad de outliers hallados en el precio de reserva de estas series es del doble que el del resto. En consecuencia, estas series pierden su carácter de cíclico, característica que vuelve al modelo autorregresivo efectivo en la predicción de valores futuros.

Además, se divisan en las series 768x1024 y 320x480 casos de extrema divergencia entre el dato procesado y el dato predicho que contribuyen a sesgar el valor del error cuadrático medio. Es decir, casos en que spline ha fallado en corregir los valores detectados como outliers. Sin estos pocos errores en los datos procesados, el error cuadrático medio se asimilaría al estimado en el resto de las series.

Sin considerar los casos anteriores, gracias a la creación del sistema de detección y tratamiento de observaciones ruidosas, se ha logrado estimar el precio de reserva con un

error cuadrático medio promedio del 0.1999601[26], lo que representa un error medio absoluto relativo al precio de reserva de 3.58%. Además, al analizar el precio de reserva por tipo y tamaño de activo individualmente, advertimos que se logra reducir del error cuadrático medio en la mitad de los casos. Sin embargo, el mayor beneficio de explorar este grado de granularidad radica en que es posible analizar la variabilidad del precio de cada serie por separado. En consecuencia, las predicciones resultantes aumentan la efectividad del análisis.

De todas formas, hemos de evaluar la posibilidad de incrementar la granularidad y/o dinamismo del modelo puesto que, en esta industria, una diferencia de pocos centavos en las estimaciones puede significar grandes costos al tratarlo en volumen, impactando en los márgenes de las DSPs.

En el análisis de los resultados, se advirtió la inexistencia de una ventana de tiempo que minimice para todas las series el error cuadrático medio. Se podría definir que este debería ser otro aspecto dinámico a tomar en cuenta en el desarrollo de un bucle más sofisticado.

Asimismo, en esta ocasión, se hizo uso de la base de datos de *Inneractive* para el mercado de Estados Unidos. Una de las particularidades no alcanzadas en este estudio es la dimensión de este país y, por lo tanto, la diferencia horaria entre regiones. Dado que el sistema desarrollado estima considerando el carácter cíclico del comportamiento diario del precio de reserva, un análisis fragmentado por zona horaria podría incrementar la eficacia del modelo. Aunque también se podría argumentar que este tipo de estudio solo resultaría de utilidad en el caso de que la campaña se focalice en ciudades específicas.

Otra posible mejora en el sistema sería la automatización de los parámetros por los que se clasifica a una observación como outlier o no. En la práctica actual, hemos definido como outlier a la observación que resulte ser 1.5 veces más grande *(0.5 más chica)* que sus valores históricos[27] y que, a su vez, su la variación porcentual con respecto a la hora anterior resulte ser mayor *(menor)* que la variación porcentual histórica[28]. Si bien este ha resultado un método eficaz, la reformulación automática de estos parámetros a través del tiempo conllevaría a una optimización del método utilizado para analizar y procesar los datos.

Por último, dado que las intuiciones bajo las que este bucle se construyó son propias del RTB, descartamos la idea de que este modelo funcione únicamente para este caso de estudio

[26] Sin considerar el error cuadrático medio de 320x480 y 768x1024.

[27] Las observaciones correspondientes al mismo horario de la observación evaluada de los 7 días anteriores.

[28] Las variaciones porcentuales entre observaciones correspondientes al mismo horario de los 7 días anteriores.

en particular. De todas formas, a modo de ejercicio sería relevante estimar para otros casos ya que posibles nuevas conclusiones podrían contribuir al perfeccionamiento el modelo.

Referencias

Adikari S., Dutta K., "Real Time Bidding in Online Digital Advertisement", *Lecture Notes in Computer Science*, Volumen 9073, B. Donnellan et. al. (Dublín, DESRIST 2015: Springer International Publishing Switzerland, 2015), pp. 19-38.

Burgers G., Jan van Leeuwen P., Evensen G., "Analysis Scheme in the Ensemble Kalman Filter", *American Meteorological Society*, Volumen 126. (Holanda: Royal Netherlands Meteorological Institute, 1998): disponible en http://journals.ametsoc.org/

Gobron N., Verstraete M. M., Musial J.P., "Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series", *Atmospheric Chemistry and Physics* (Italy: European Commission, Joint Research Centre, Institute for Environment and Sustainability, 2011): disponible en https://www.atmos-chem-phys.net/11/7905/2011/

Google, "The Arrival of Real-Time Bidding and What It Means for Media Buyers", (2011) Google Inc.: disponible en https://static.googleusercontent.com/media/www.google.com/en//doubleclick/pdfs/Google-White-Paper-The-Arrival-of-Real-Time-Bidding-July-2011.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R., *An introduction to statistical learning*, Volumen 6, (New York: Springer, 2013).

Lee K., Jalali A., Dasdan A., "Real Time Bid Optimization with Smooth Budget Delivery in Online Advertising", *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, KDD '13. (2013) ed. ACM: disponible en http://www.kdd.org/kdd2013/

IAB, "Q3 2016 Internet Ad Revenues Hit $17.6 Billion, Climbing 20% Year-Over-Year, According to IAB", *Internet Advertising Revenue Report,* (2016 [citado el 13 de abril 2017]): disponible en https://www.iab.com/

IAB & PwC, "IAB internet advertising revenue report, 2015 full year results April 2016", *Internet Advertising Revenue Report,* (2016 [citado el 13 de abril 2017]): disponible en https://www.iab.com/

IAB & PwC, “IAB internet advertising revenue report, 2016 full year results April 2017”, *Internet Advertising Revenue Report*, (2016 [citado el 13 de noviembre 2017]): disponible en https://www.iab.com/

Muthukrishnan, S., “Internet Ad Auctions: Insights and Directions”, *Automata, Languages and Programming*, Parte 1, L. Aceto et al. (Reikiavik, ICALP 2008: Springer-Verlag Berlin Heidelberg, 2008), pp. 14-23.

Ramamritham K., Stankovic J. A., “Scheduling Algorithms and Operating Systems Support for Real-Time Systems” *Proceedings of the IEEE*, Volumen 82, N1 (1994): disponible en https://ieeexplore.ieee.org/abstract/document/259426/

Sosa-Escudero, W. “Componentes principales y dimensionalidad”, presentación en PDF (2016), Universidad de San Andrés.

Wu W. C., Yeh M., Chen M., “Predicting Winning Price in Real Time Bidding with Censored Data”, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* KDD '15: 1305-1314. (2015) ed. ACM: disponible en http://www.kdd.org/kdd2015/

Xiao B., Yang W., Li J., “Optimal Reserve Price for the Generalized Second-Price Auction in Sponsored Search Advertising”, *Journal of Economic Commerce Research*, Volumen 10 N3 (2009 [citado el 16 de abril 2017]): disponible en http://www.jecr.org/node/108

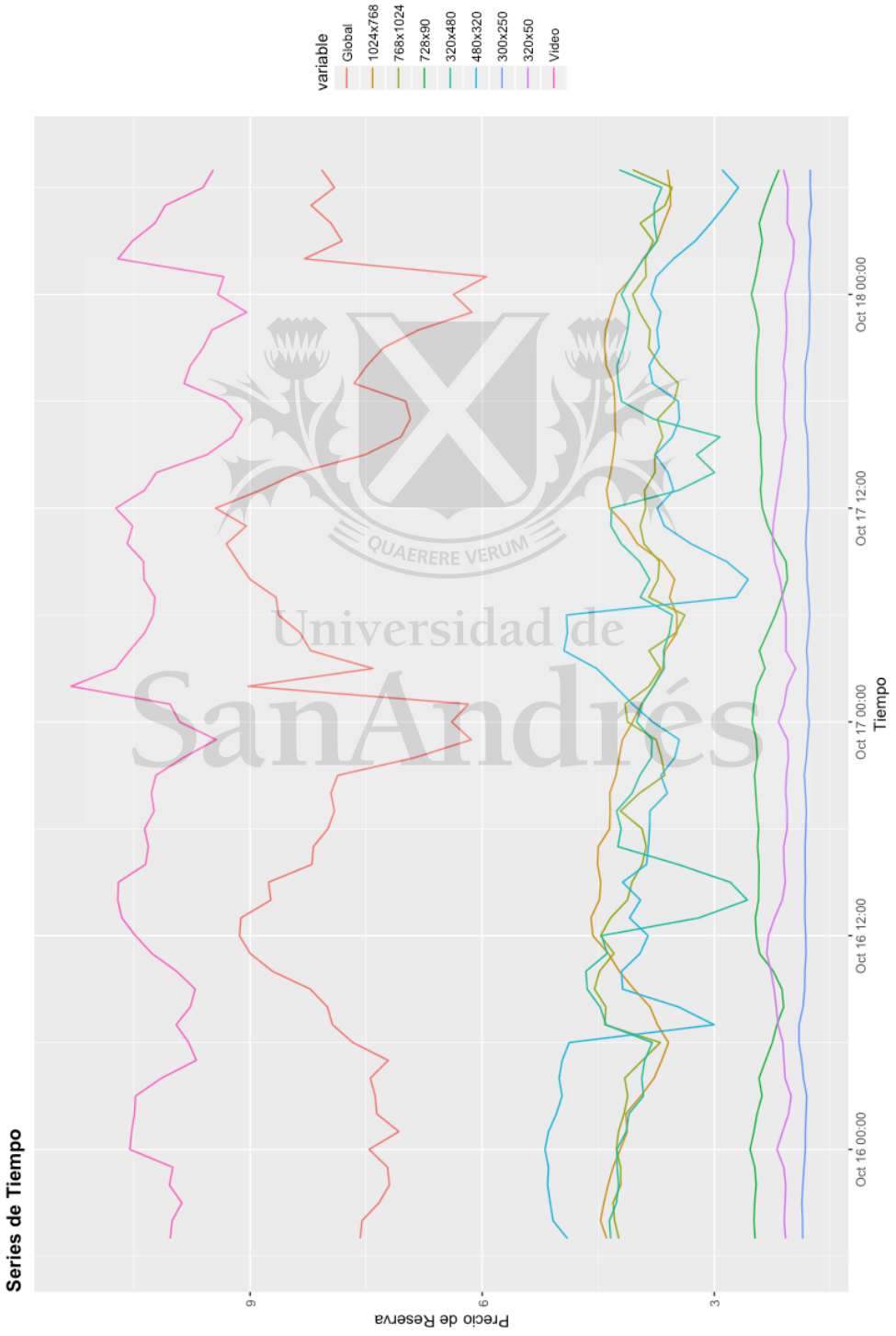
Yuan S., Wang J., Zhao X., “Real-time Bidding for Online Advertising: Measurement and Analysis”, *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, KDD '13, Article No. 1. (2013) ed. ACM: disponible en http://www.kdd.org/kdd2013/

Yuan S., Wang J., Chen B., Mason P., Seljan S., “An Empirical Study of Reserve Price Optimisation in Real-Time Bidding”, *Proceedings of the 20th ACM SIGKDD International conference on Knowledge discovery and data mining*, KDD '14: 1897-1906. (2014) ed. ACM: disponible en http://www.kdd.org/kdd2014/

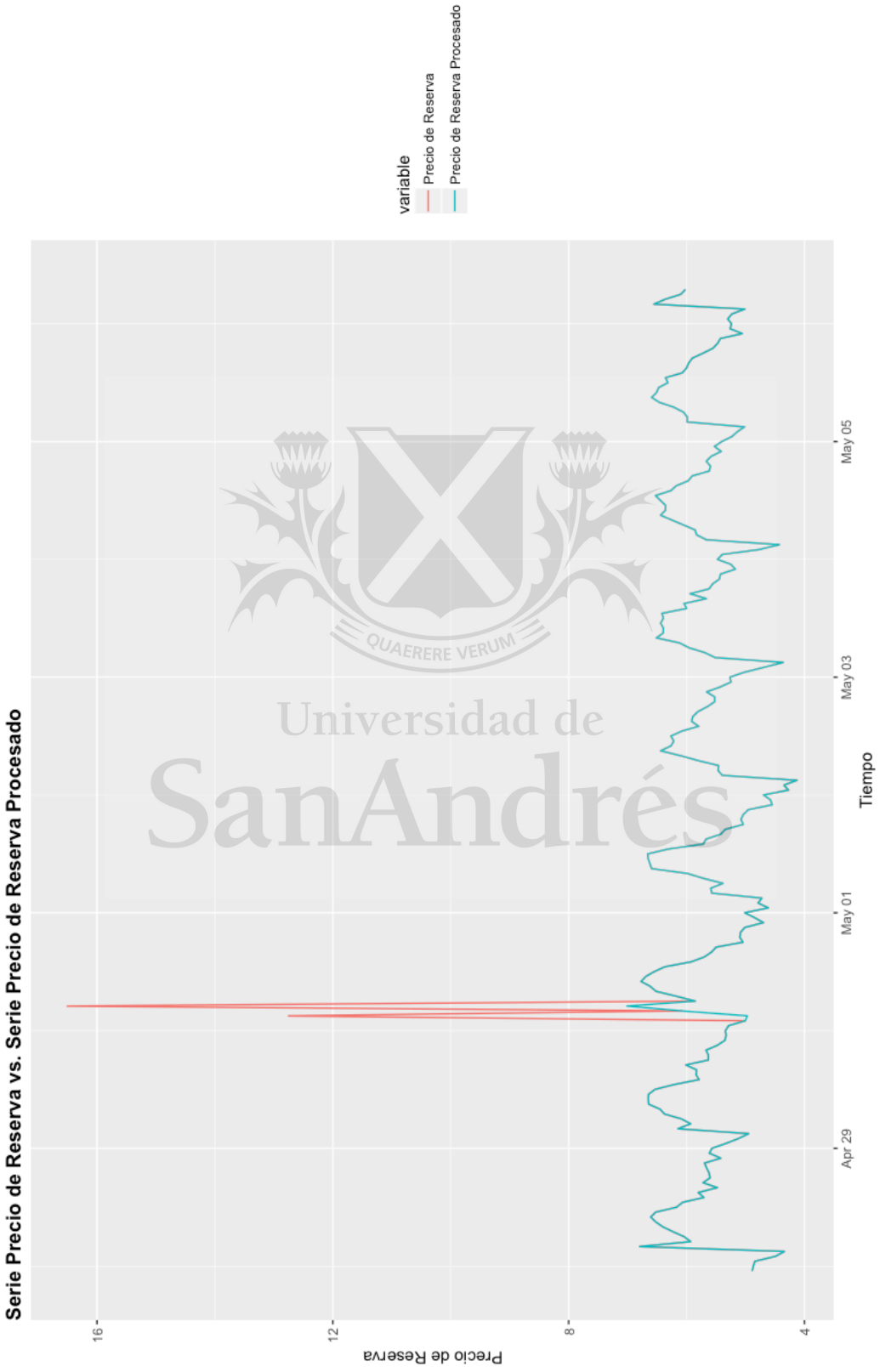
Zhang W., Yuan S., Wang J., “Optimal Real-Time Bidding for Display Advertising”, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining,* KDD '14: 1077-1086. (2014) ed. ACM: disponible en http://www.kdd.org/kdd2015/

Anexo

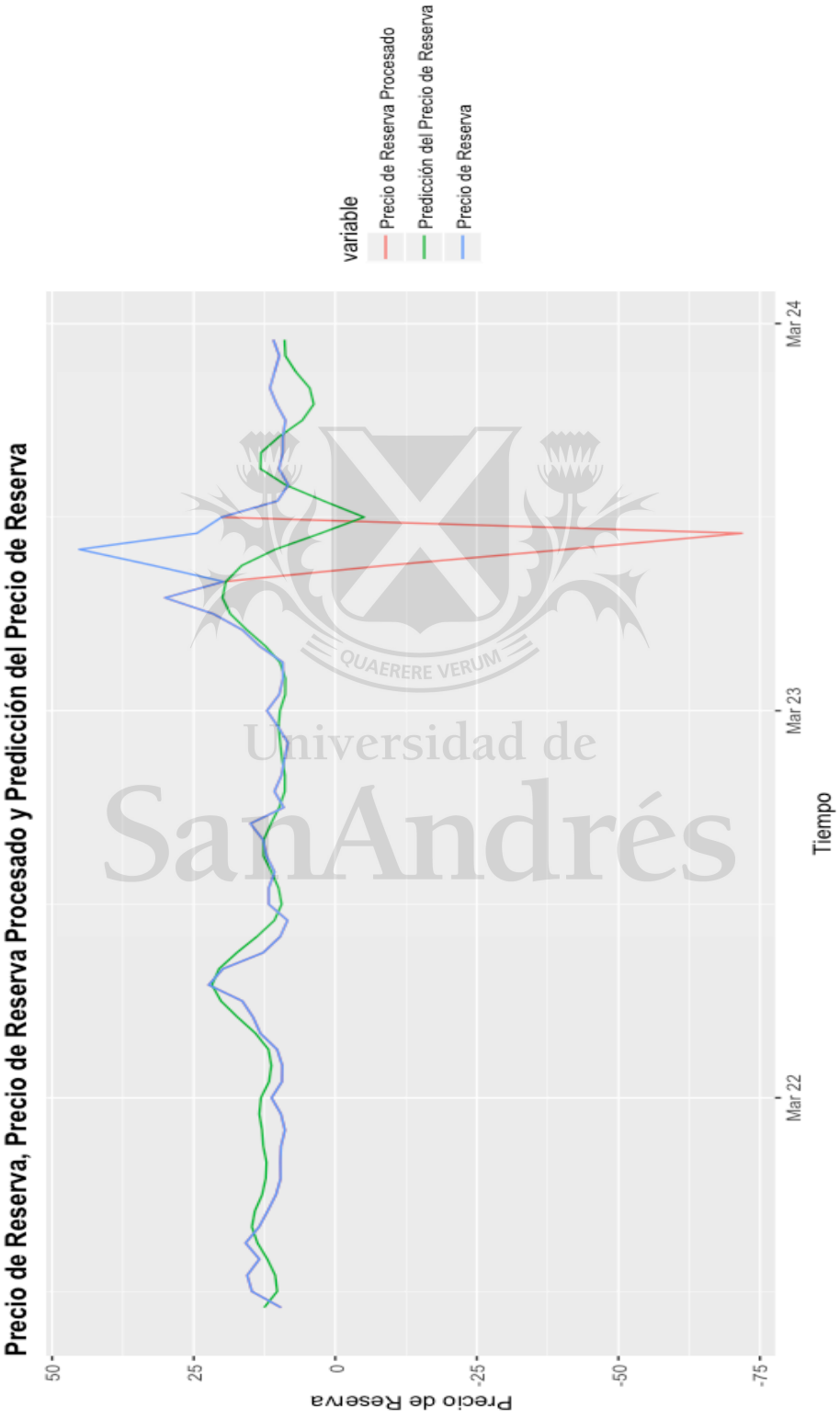
Anexo I: Series de tiempo. Precio de reserva promedio desglosado por tamaño y/o tipo de activo



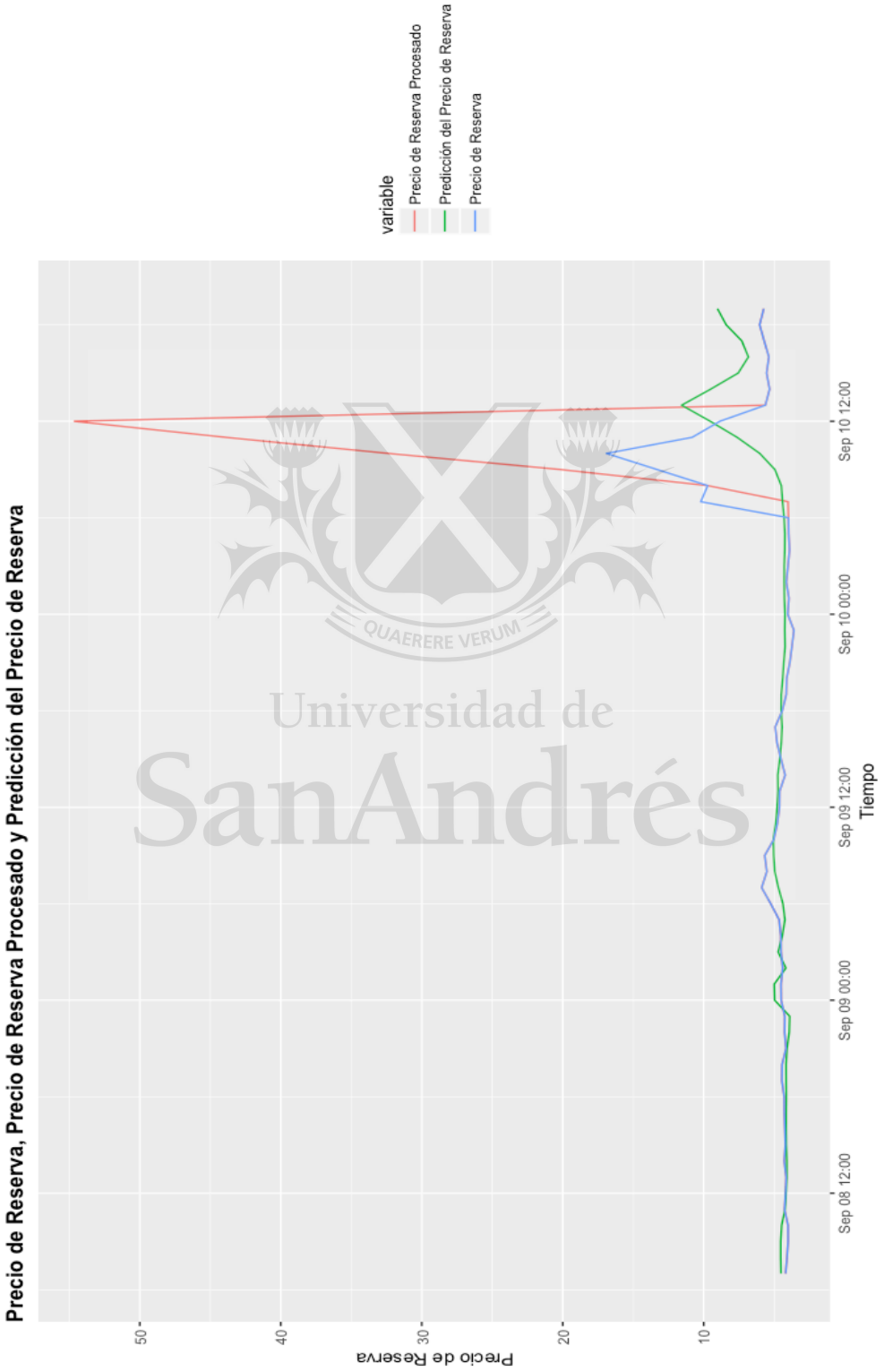
Anexo II: Datos vs. datos procesados



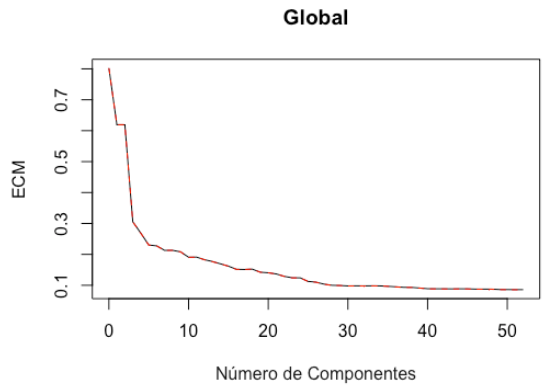
Anexo III: Precio de reserva. Datos, datos procesados y predicción – Serie 320x480

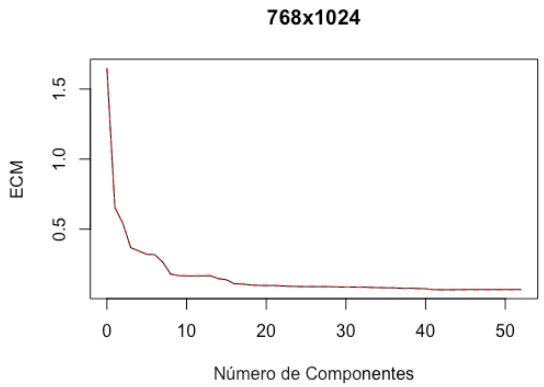


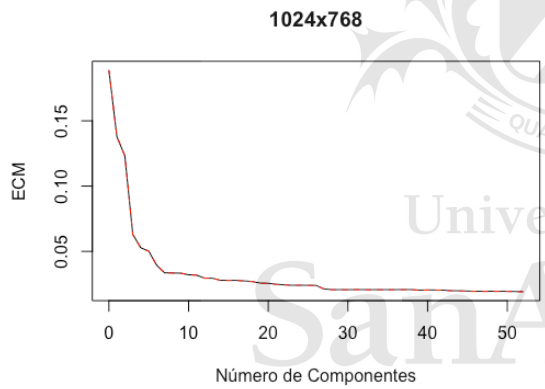
Anexo IV: Precio de reserva. Datos, datos procesados y predicción – Serie 768x1024

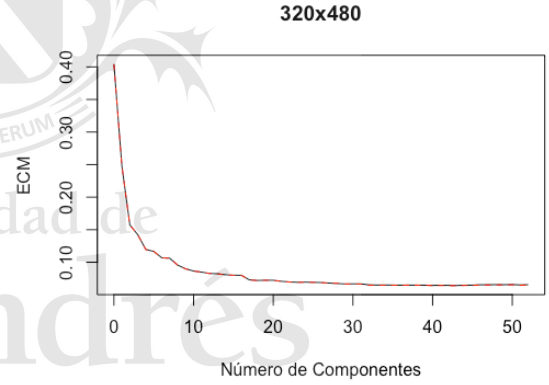


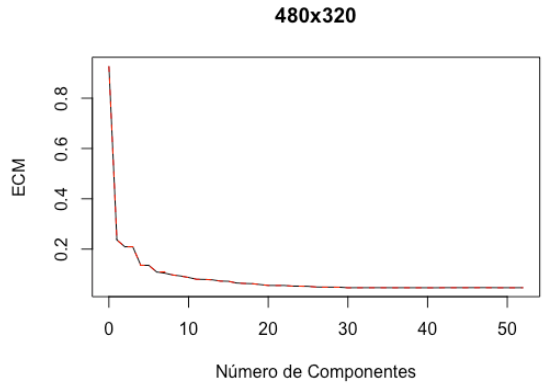
Anexo V: Error cuadrático medio de la predicción por tamaño/tipo de activo

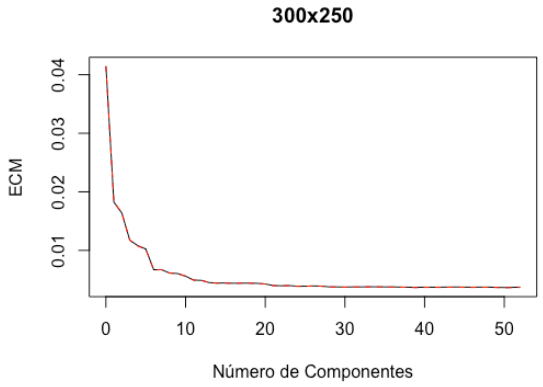


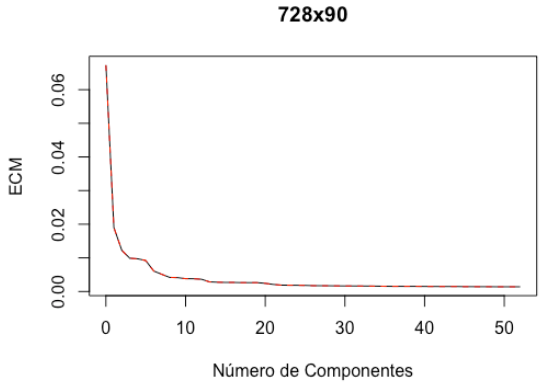


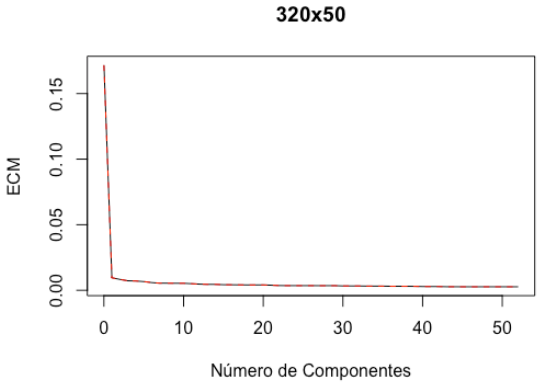


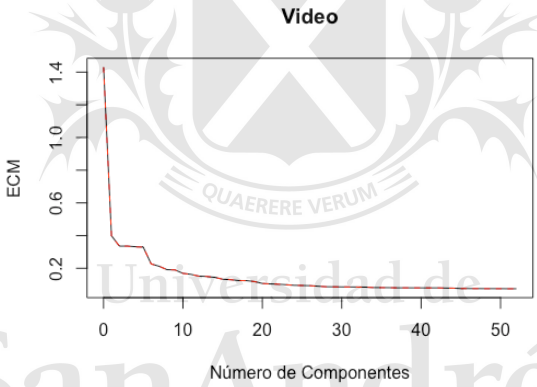












Anexo VI: Estructura de los componentes principales – Serie global