Estudio métrico sobre la actividad investigadora usando el software libre R: el caso del sistema universitario gallego

Javier Tarrío-Saavedra*
Elena Orois**
Salvador Naya*

Artículo recibido:
28 de mayo de 2015
Artículo aceptado:
28 de octubre de 2016

Resumen

Este trabajo representa una nueva alternativa para el estudio, clasificación y comparación de la producción científica de centros de investigación, utilizando las funciones de tratamiento de datos del paquete Citan del software estadístico R. En particular, se muestra el análisis bibliométrico de las publicaciones de las universidades de A Coruña, Santiago de Compostela y Vigo, en el periodo 2000-2011, recopiladas por la base de datos Scopus. Entre las técnicas usadas se aplicaron modelos de Lotka y Price, modelización no paramétrica y paramétrica de los datos, así como el cálculo y

* Escola Politécnica Superior, Universidade da Coruña. javier.tarrio@udc.es, salva@udc.es.
** Servicio de Biblioteca, Universidade da Coruña. elena.orois@udc.es.

análisis de indicadores de la cantidad y calidad de la producción científica, los índices h y g, y otros menos conocidos como los rp_1, l_p, Ge_1, Ge_5 y Sl_{p1}. Como novedad, se propone una variante del índice h (h_h) que define el grupo de investigadores que forman la élite más productiva de cada universidad y estima su calidad investigadora.

Palabras clave: Bibliometría; Cienciometría; Índice h; Índice g; Scopus; Software R; Comunicación científica.

Metric study of the research activities using free software R: The case of Galician university system

Javier Tarrío-Saavedra, Elena Orois and Salvador Naya

Abstract

This work represents a new alternative for the study, classification and comparison of the scientific production corresponding to research entities. It consists on the application of statistical data processing functions available in the R software's Citan package. In particular, the bibliometric study of publications of universities of A Coruña, Santiago de Compostela and Vigo, in the period 2000-2011, compiled by the Scopus database. The study was conducted using the statistical analysis of the data, the application of models of Lotka and Price, nonparametric and parametric modeling (Pareto) of the data, and the calculation and analysis of indicators of the scientific production like the h and g indexes, and others lesser known as rp_1, l_p, Ge_1, Ge_5 and Sl_{p1}. A novelty consists in a variant of the h index (hh) that defines the group of researchers who are the most productive of each university, the elite, and estimates the researching quality of such representative elites.

Keywords: Bibliometrics; Scientometrics; h index; g index; Scopus; Software R; Science communication.

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

Introducción

La necesidad de conocer y evaluar con precisión la actividad investigadora de las universidades es un problema que ocupa y preocupa no sólo a las instituciones involucradas, sino también al resto de la sociedad, en tanto son instituciones públicas. En este estudio se propone el uso del software libre R para analizar la producción científica de las tres universidades gallegas.

Tradicionalmente, dentro de las ciencias de la documentación se diferencian tres tipos de disciplinas dedicadas a la métrica de la información: Bibliometría, Cienciometría e Informetría. Estas tres herramientas del tratamiento de la información documental presentan características comunes y cierto solapamiento en relación con el flujo de información y a los modelos que utilizan. En muchos trabajos cuyo objeto es la literatura científica, la Bibliometría, la Cienciometría y la Informetría aparecen situadas al mismo nivel, llegándose en algunos casos al extremo de considerar estos conceptos como sinónimos, dificultando la identificación de sus límites y alcance respectivos. Aun así, es posible encontrar algunas definiciones que permiten situar las fronteras entre los mencionados conceptos.

El primero en hablar de Bibliometría fue Pritchard (1969), quien la concibe como "la aplicación de las matemáticas y los métodos estadísticos a los libros y otros medios de comunicación". Por otra parte, Brookes (1990), quien realizó importantes aportaciones a la idea de la obsolescencia de la literatura científica, él entiende la Bibliometría como "una disciplina limitada a la actividad bibliotecaria que debe enriquecerse mediante las relaciones interdisciplinarias con la estadística para refinar sus técnicas". En este sentido, la Bibliometría tendría por objeto analizar libros y revistas científicas y, principalmente, comprender las actividades de comunicación de la información. Quizás la mejor definición sea la que aporta Alcaín (1991), quien enuncia que la Bibliometría es la disciplina que "comprende estudios dirigidos a conocer el rendimiento de los fondos de publicaciones científicas, así como la selección y el consumo por parte de los usuarios".

Por otra parte, las primeras menciones a la Cienciometría se encuentran en un trabajo de Hulme (1923), en el que se realiza un análisis estadístico de la historia de las ciencias. Sin embargo, la popularidad de esta disciplina surgió a partir de 1977, con el nacimiento de la revista *Scientometrics*. La Cienciometría alcanzó un gran desarrollo con los trabajos de Eugene Garfield (1995). A partir de la aparición de la obra de Price en 1963 (*Little Science, Big Science*[1973]), nace y se desarrolla un nuevo campo en las ciencias documentales: el análisis estadístico y sociométrico de la literatura científica, denominado Cienciometría "cuando los estudios se realizan sobre la

producción y productividad de los autores científicos y organismos de investigación". Una definición tentativa de Cienciometría podría ser "ciencia de la ciencia".

El término Informetría comienza a utilizase a partir de los años ochenta. Spinak (1996) enuncia que la Informetría se basa en las investigaciones de la Bibliometría y la Cienciometría, comprendiendo tareas como el desarrollo de modelos teóricos para la medida de la información. Mediante su aplicación, se busca encontrar, definir y caracterizar regularidades, patrones, asociaciones en los datos relacionados con la producción y el uso de la información registrada. Además de la medición de la información, la Informetría también se ocupa de su almacenamiento y recuperación; por lo tanto, diríase que es una disciplina instrumental de las ciencias de la documentación, que se apoya en el empleo de las herramientas informáticas.

Posteriormente, en los noventa, con la aparición y el desarrollo de la Internet como instrumento de comunicación científica, han surgido estudios cuantitativos cuyo objeto es la información electrónica contenida en el ciberespacio: Cibermetría y Webmetría. De nuevo nos encontramos ante términos utilizados como sinónimos, si bien el campo de la Cibermetría se considera más amplio, ya que comprende el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías en la Internet, desde perspectivas bibliométricas e informétricas.

Mientras que la Webmetría se circunscribe a portales de información o webs concretas, como el estudio métrico de la información de una web en particular. En estos nuevos análisis, el impacto que suponían las citas en el formato impreso se sustituye básicamente por la medición de la cantidad y calidad de los enlaces presentes en la web.

Finalmente, en el 2010 apareció el concepto de Altmetrics, como la creación y el análisis de nuevas métricas basadas en la web social, para analizar y evaluar a los investigadores. Aquí los nuevos indicadores son las valoraciones que las contribuciones reciben por parte de otros usuarios en la web social, por ejemplo, los "me gusta" de Facebook.

Entre los indicadores más utilizados actualmente en el campo de la Bibliometría están dos muy importantes que se calculan y analizan en este artículo: el índice h y el índice g. El primero fue propuesto por Jorge E. Hirsch (2005) y se define como un número H que coincide con mayor número de publicaciones de ese autor que recibieron al menos H citas. Un índice H = X significa que hay X artículos con X o más citas, pero no X+1 que tengan X+1 o más citas (Arencibia y Carvajal, 2008).

Al igual que con el cálculo del índice h, para obtener el índice g se listan los artículos de un autor en orden descendente, de acuerdo con el número

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

de citas recibidas. El mayor número de orden en el listado, donde la suma de las citas recibidas por el autor sea mayor o igual al cuadrado del número de orden, será considerado el índice G de dicho autor (Arencibia y Carvajal, 2008).

El índice g fue creado por Leo Egghe (2006) y se define como el (único) número G más grande, de tal manera que los artículos menores que G recibieron (en conjunto) al menos G2 citas. El índice g mantiene todas las propiedades positivas del índice h y las mejora, al contemplar todas las citas de los artículos más citados, propiedad que permite distinguir entre científicos con índices h muy similares, pues el número g es mayor y más variable que el h.

Aparte de los índices h y g, en los últimos años se han desarrollado multitud de nuevos números índice de agregación. En este trabajo se incluyen los índices SL_{p1} (estadístico cuasi L), el l_p (ideal para evaluar a productores de literatura científica todavía con escasa producción), los índices Ge_1 (número de artículos con al menos una cita) y Ge_5 (número de artículos con al menos cinco citas) y, por último el r_p (generalización del índice h y g) . Además de estos últimos, todo análisis bibliométrico debería completarse con el cálculo de otros índices más tradicionales, como el número total de documentos producidos, el número máximo de citas a un documento producido y el número total de citas recibidas al conjunto de todos los documentos relativos al centro de investigación estudiado. Seguidamente se muestran las expresiones matemáticas de las diversas funciones de impacto introducidas, en las cuales x es el número de citas de cada documento producido. Para más información consúltese el trabajo de Gagolewski (2011).

$$h(x) = \max\{i = 0,1,\dots,n: x_{(n-i+1)} \geq i\}/X \in \mathbb{N}_0^n$$

$$g(x) = \max\left\{i = 0,1,\dots,n: \sum_{k=1}^{i} x_{(n-i+1)} \geq i^2\right\}/X \in \mathbb{N}_0^n$$

$$Ge_1 = \sum_j \mathbb{I}_{\{x_j \geq 1\}} \;;\; Ge_5 = \sum_j \mathbb{I}_{\{x_j \geq 5\}} \;;\; SL_{p1} = \sum_j \ln(x_j + 1) \;;\; l_p = \operatorname{argsup}\{ab: e^{p,a,b}\}$$

$$r_p(x,p) \coloneqq \sup\{r > 0: s^{p,r} \trianglelefteq x\} \vee X \in \mathbb{N}_0^n \; y \; s^{p,r} \in \mathbb{N}_0^{\lceil r \rceil} \; con \; r > 0, siendo$$

$$s^{p,r} = \begin{cases} \sqrt[p]{r^p - 0^p}, \sqrt[p]{r^p - 1^p}, \dots \sqrt[p]{r^p - \lceil r - 1 \rceil^p} & si \; p < \infty \\ (r, r, \dots, r) & si \; p = \infty \end{cases}$$

En este trabajo se incluye la modelización de la producción científica recopilada en la base de datos Scopus, para lo cual se aplican y evalúan los conocidos modelos de Price (1973) y Lotka (1926). La Ley de Price de crecimiento exponencial enuncia que la ciencia crece a interés compuesto,

multiplicándose por una cantidad determinada en periodos iguales de tiempo, donde *N* es el número de publicaciones en un periodo determinado, *t* es el tiempo y *b* es un parámetro del modelo relacionado con la velocidad de crecimiento. Respecto de la ley de Lotka, es el modelo de regresión no lineal que relaciona el número de autores con su productividad: "El número de autores, A_n, que publican *n* trabajos sobre una materia es inversamente proporcional al número de artículos al cuadrado". Implica que muy pocos autores publican la mayoría de los trabajos. El valor de los parámetros pertenecientes a los dos modelos mencionados permite no sólo definir la relación entre variables bibliométricas, sino que, de la misma forma, aporta una información muy valiosa acerca de la producción científica en las entidades de investigación analizadas, haciendo la comparación entre sí.

Las tres universidades aquí examinadas desde un punto de vista bibliométrico son las que conforman el Sistema Universitario de Galicia (SUG): la Universidad de Santiago de Compostela (USC), la Universidad de A Coruña (UDC) y la Universidad de Vigo (UVigo), siendo las tres instituciones de referencia representativas de cada ciudad donde se ubican, poseen una historia y circunstancias diferentes, en particular si se compara la USC respecto de las demás. De hecho, esta universidad se creó en 1495, lo que la convierte en una de las más antiguas de España, siendo una universidad de amplia tradición institucional y formativa, con titulaciones fuertemente implantadas y reconocidas, como Medicina o Derecho.

Por otro lado, las universidades de A Coruña y Vigo se fundaron hacia 1990 mediante la Ley 11/1989, de 20 de julio, de ordenación del sug. Pero, además de su relativa juventud, éstas también tienen en común el estar ubicadas en ciudades industriales y comerciales portuarias, de similar tamaño y estructura (no son capital de comunidad autónoma).

En cualquier caso, las tres instituciones de educación superior reciben la mayor parte de su financiación de la Xunta de Galicia, en función de unos indicadores preestablecidos, de ahí la importancia de conocer la productividad de esas universidades para valorar su contribución al desarrollo de la ciencia y, por tanto, de la comunidad donde están y, en definitiva, la eficiencia de la inversión pública realizada.

Así, mediante la descripción, comparación, modelización y evaluación estadística de la actividad investigadora de las tres universidades gallegas y sus colaboradores durante el periodo 2000-2011, intentamos conocer y evaluar de forma precisa su producción científica, analizar sus problemas y detectar sus ventajas competitivas. La metodología aquí propuesta puede aplicarse perfectamente a los datos de producción de cualquier otro centro universitario, de un modo rápido y eficiente.

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

Este trabajo está estructurado como se indica: en el segundo apartado se presenta una breve descripción de las herramientas y la metodología definidas para el estudio bibliométrico comparativo, en el tercer apartado se muestran y comentan los resultados obtenidos a partir de su aplicación y, finalmente, en el cuarto acápite se enuncian las conclusiones principales y recomendaciones.

Herramientas y metodología

Como ya dijo, la necesidad de conocer y evaluar de manera precisa la actividad investigadora en estas tres universidades gallegas, analizar sus debilidades y detectar sus fortalezas obliga al uso de herramientas estadísticas. El objetivo básico de este trabajo es describir, comparar, modelizar y evaluar estadísticamente la actividad investigadora de las tres universidades gallegas y sus colaboradores en el periodo 2000-2011, mediante la aplicación de paquetes estadísticos pertenecientes al software libre R. En particular, mediante el uso de la librería CITation ANalysis toolpack (Citan), desarrollada por Gagolewski. La metodología empleada se muestra a continuación.

Obtención de datos

Para realizar el análisis, es necesario contar con una base de datos bibliográfica que recoja de forma fiable la producción de los centros de investigación en cuestión. SciVerse Elsevier Scopus es la base de datos bibliográfica general escogida, ya que comprende un mayor número de fuentes o publicaciones con sistema de revisión entre pares. De hecho, según la Academic Database Assessment Tool (ADAT), el 1º de enero de 2014, Scopus indexaba veinte mil revistas científicas con revisión entre pares y 370 "book series", pertenecientes a cinco mil editores diferentes (Elsevier, Springer-Verlag, Kluwer Academic Publishers, entre otros) (Center for Research Libraries, 2014).

Cada fuente de Scopus está caracterizada con cuatro dígitos All Science: Journal Classification (ASJC) que determinan su temática o categoría principal. Así, la secuencia 1,200 corresponde a "Arts and Humanities", la 1,600 a "Chemistry", y así por el estilo.

En la base de datos Scopus, cada referencia está definida por una serie de campos, catorce de los cuales son reconocidos y utilizados por el paquete Citan del software estadístico R: Authors, Title, Year, Source Title, Volume, Issue, ArticleNumber, PageStart, PageEnd, Citations, UniqueId, issn, Language and DocumentType.

Scopus utiliza la siguiente clasificación de tipología documental, que posteriormente se empleará para resumir y analizar estadísticamente la producción científica: "Article" (ar), "Conference paper" (cp), "Review" (re), "Short survey" (sh), "Article-in-Press" (ip), "Note" (no), "Erratum" (er), "Editorial" (ed), "Letter" (le).

Esta base de datos ofrece una serie de aplicaciones para el tratamiento estadístico descriptivo de los datos bibliográficos, pero, al igual que ocurre con otras bases de datos, como la Web of Science, el manejo de la información está limitado a un número máximo de referencias. En el caso particular de Scopus, aunque hoy se exportan hasta un máximo de vinte mil registros en formato .csv, cuando realizamos este artículo el tamaño máximo se limitaba a dos mil referencias, lo que obligó a exportarlos en bloques de dos mil registros y posteriormente unirlos para obtener una única base de datos con las publicaciones de las tres instituciones educativas.

La posibilidad de exportar registros en formato .csv permite el análisis de la producción científica de centros de investigación de gran magnitud, como es el caso de las universidades o de los países. De ahí el uso de un programa estadístico externo que permita la formación de la base de datos requerida y su posterior tratamiento estadístico está totalmente justificado, atendiendo a las necesidades actuales de control y evaluación de la rentabilidad de la producción científica.

La búsqueda en Scopus se realizó indicando todas las posibles variables en los nombres de los centros. En este caso, fue relativamente sencillo obtener un conjunto de datos fiable de la USC y UVigo, mas no así en el caso de la UDC, debido a las variantes ortográficas del nombre en las diferentes lenguas utilizadas por los autores (gallego, inglés y español) y a la carencia de un criterio único para utilizar el nombre de la institución de forma normalizada.

Cabe destacar que la información recogida, en particular la que alude al número de citas de cada documento, se obtuvo a partir del 2 de junio de 2012. La elección de este periodo se debe a la falta de fiabilidad en los datos de las referencias más recientes, pues normalmente los índices de impacto, como el jcr, o la misma incorporación de publicaciones a las bases de datos se hizo con un cierto retraso de un año, por lo que los datos del 2013 estarían disponibles a mediados del 2014.

Tratamiento estadístico de los datos

El uso de las herramientas estadísticas es absolutamente necesario en el tratamiento de información bibliométrica. Al respecto, R es un software libre,

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

considerado como la *lingua franca* de la estadística computacional, con la ventaja de disponer de la más completa oferta de librerías para la aplicación del análisis estadístico de datos: regresión lineal y no lineal, paramétrica y no paramétrica, análisis de la varianza, estadística espacial, análisis de datos funcionales, series de tiempo, reducción de dimensión, reconocimiento de patrones, minería de datos, entre otros.

Como ya se comentó, en el caso del análisis bibliométrico se recurrió a la librería Citan, usando la base de datos Scopus. Seguidamente se muestra la secuencia comprendida entre la instalación del paquete Citan y la obtención de la base de datos, a partir de la cual se realizó el estudio bibliométrico:

- Se instala el paquete Citan en R, mediante la función install.packages().
- Seguidamente se carga Citan, library(Citan).
- Se crea una conexión con una base de datos llamada Baseudc.db: conn<-lbsConnect("Baseudc.db").
- Se crea el esquema o base de datos relacional: lbsCreate(conn).
- Se importan las fuentes Scopus: Scopus_ImportSources(conn).
- Se van importando los datos de cada consulta en formato csv: data <- Scopus_ReadCSV("UDC _New_1.csv").
- Y se adjuntan en una base de datos relacional llamada Resultsudc: lbsImportDocuments(conn, data, "Resultsudc").

Resultados y discusión

Estudio estadístico descriptivo

Una vez obtenida una base de datos fiable, correspondiente a cada universidad durante el periodo 2000-2011, se procede a realizar el análisis estadístico de su producción científica, mediante el uso de las funciones pertenecientes al paquete Citan de R. El primer paso en un análisis de estas características es el estudio descriptivo de los datos. La función lbsDescriptive.

Stats(conn,surveyDescription="ResultsUDC") de la librería Citan proporciona un completo y útil punto de partida para el análisis bibliométrico, obteniéndose resultados vinculados con la producción científica, según el tipo de documentos, producción de los autores, categoría de documentos y antigüedad. La aplicación a los datos de cada una de las universidades proporciona una serie de salidas en forma de cuadros y gráficos que se muestran a continuación y que, convenientemente modificadas, sirven para comparar la producción de las tres instituciones.

Los primeros resultados obtenidos se refieren a la producción total de los centros en el periodo estudiado, es decir, el número de documentos donde al menos uno de los autores está afiliado a la universidad examinada. Además, esta primera salida proporciona el número de autores participantes, obviamente no sólo los que presentan afiliación al centro educativo en cuestión, sino el número total de autores que elaboraron los documentos analizados, la red de investigación completa, incluyendo los colegios invisibles de cada universidad gallega.

Concretamente, en el caso de la USC, los artículos obtenidos a partir de la base de datos Scopus fueron 10,095, cuya autoría corresponde a un total de 19,492 firmantes, lo que proporciona una ratio de 0.19 documentos por autor. En cuanto a la UDC, los documentos analizados y el número de autores ascienden a 4,080 y 6,206, respectivamente, resultando una ratio de 0.66 documentos por autor.

Finalmente, en lo que respecta a la UVigo, se identificó un total de 8,230 documentos, correspondientes a 10,817 autores (0.76 documentos por autor). Por lo tanto, es evidente que la universidad que genera más documentos es la USC (de hecho, duplica a las demás), aunque las más productivas son las más recientes, es decir, la UVigo y la USC, por este orden y a una distancia relativa de la primera.

A esta altura, conviene destacar las dificultades que, para obtener una base de datos completa y fiable, origina el uso por parte de los autores de distintas denominaciones para indicar su afiliación; este problema es especialmente acuciante en el caso de la Universidad de A Coruña. El uso de la denominación oficial traducida a distintas lenguas (español, gallego, inglés, francés…), la sustitución de la "ñ" por la "n", la omisión del artículo, incluso la sustitución de toda referencia a la universidad, empleando en su lugar el nombre del campus (Ferrol, por ejemplo) o del centro de trabajo, dificultan la recuperación de la información.

En definitiva, el uso de la denominación oficial de cada universidad debería ser norma para evaluar con mayor fiabilidad y eficiencia su producción científica, previniendo la pérdida de tiempo y recursos relacionados con el filtrado y tratamiento previo de la base de datos.

Una vez contabilizados los documentos y autores pertenecientes a la red de investigación de cada universidad gallega, el siguiente paso fue calcular y describir su distribución de la producción científica, según la categoría del documento dentro de la clasificación que utiliza Scopus (Física y Astronomía, Matemáticas, Medicina, etc.). Esto se hace de un modo sencillo y claro, mediante la construcción de diagramas de sectores, en los que cada campo representa una categoría Scopus o campo de la ciencia. La *Figura 1* muestra

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

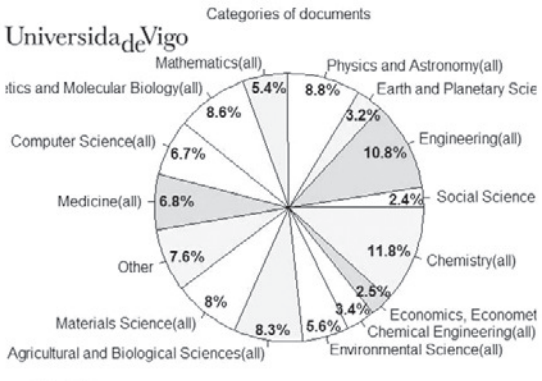
los diagramas de sectores para cada universidad, en donde se observa la distribución de su producción científica por categoría, en porcentajes.

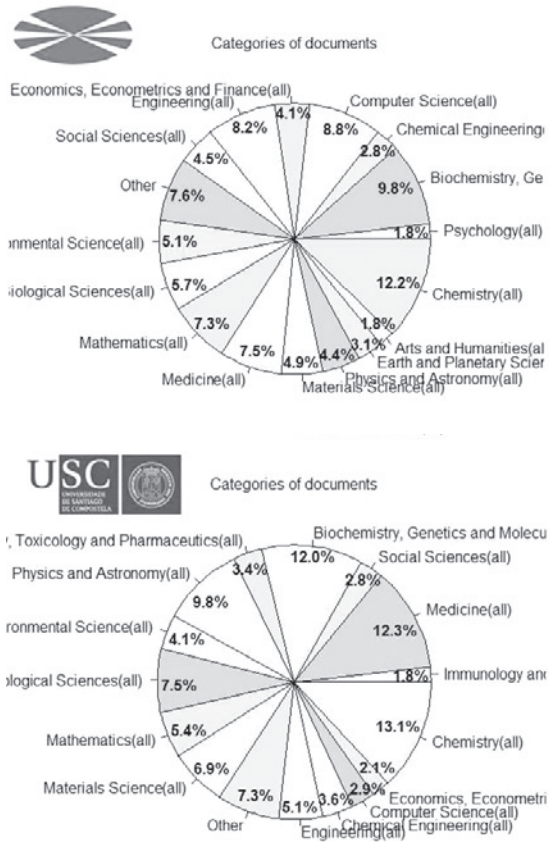
Destaca que "Chemistry" (Química) es la categoría con más relevancia, o lo que es lo mismo, con mayor producción, en las tres universidades; más importante resulta si se estudia en conjunto con la categoría "Biochemistry, Genetics and Molecular Biology" (Bioquímica, Genética y Biología Molecular). Esto no es en absoluto un hecho aislado o un resultado sorprendente, pues es bien conocido que hay categorías con mayor producción que otras, ya sea por el número de fuentes existentes, número de investigadores, condicionantes políticos y económicos, etc., siendo la química, en general, una de las categorías más productivas.

Aparte de esta última circunstancia, se observa la existencia de una serie de categorías "franquicia", características de cada universidad, con una importancia mayúscula en su producción. Éstas son Engineering (Ingeniería) en la UVigo, Medicine (Medicina) en la USC y Computer Science (Informática) en la UDC. Esta importancia relativa deriva de la existencia de facultades y escuelas únicas en cada universidad, marcando fuertemente el carácter de cada cual: las ciencias de la vida en la USC (médico, químico, bioquímico), un perfil más ingenieril en la UDC y, sobre todo, en la UVigo.

Otro resultado interesante es el hecho del poco peso que las humanidades y las ciencias sociales ocupan en el sistema universitario gallego, en comparación con las restantes ramas. Para ilustrar las salidas propias de la librería Citan de la forma más fiel posible, en lo sucesivo los gráficos resultantes se muestran en su versión original, con leyendas y textos en inglés.

Figura 1. Diagrama de sectores: distribución de los documentos, según categoría Scopus o campo de la ciencia, para la UVigo, UDC y USC



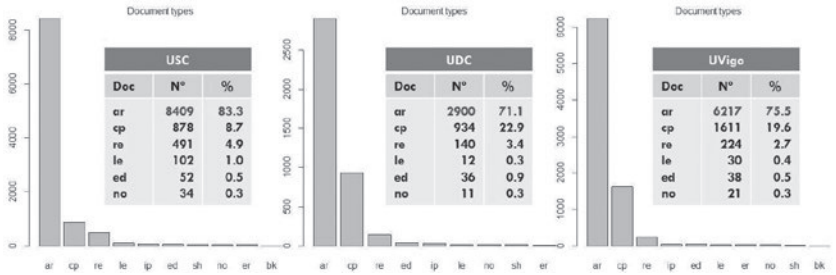


Aparte de la distribución de los documentos según sus categorías, para una completa comparación entre centros, también es importante caracterizar su distribución respecto del tipo de documento. La *Figura 2* es un diagrama de barras que muestra este tipo de distribución para las tres universidades gallegas; así, en ordenadas se muestra la frecuencia o número de documentos, y en abscisas el tipo de documentos.

Además, con el objeto de aportar una información cuantitativa más precisa, se ha modificado el gráfico de salida Citan, incluyendo tablas de frecuencia absoluta y relativa (en porcentaje) para cada universidad. Los resultados contenidos en la *Figura 2* muestran que, en general, el tipo de documento más frecuente, con diferencia, es el artículo de revista o "paper" (ar), seguido, a cierta distancia, del artículo de congreso o "conference paper" (cp). Obviamente, escribir y enviar un artículo es más económico que asistir a un congreso; además, la posibilidad de acceder a una publicación cp no es siempre factible.

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

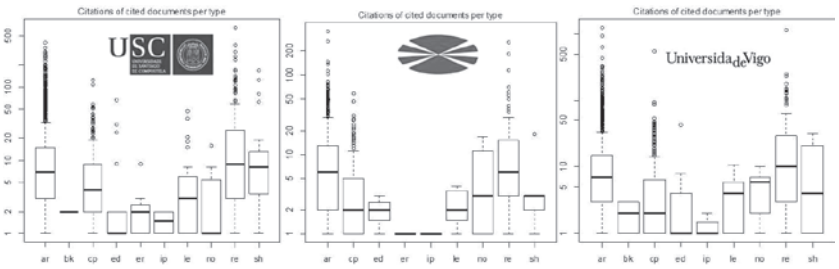
Figura 2. Diagramas de barras: distribución de los documentos, según su tipo



Pero cabe destacar que la proporción de documentos cp es mucho mayor en la UDC y UVigo. Se observa una posible causa en el peso que la ingeniería y la informática desempeñan en estas dos universidades, categorías estrechamente vinculadas a la oferta de una gran cantidad y variedad de congresos científicos, que además presentan una mayor tradición en lo que respecta a la edición y publicación de "proceedings" y números especiales en revistas del campo.

Otro resultado interesante se obtiene cuando, precisamente respecto del tipo de documento, se calcula el número de citas. El resultado se observa en la *Figura 3.* Como era de esperar, ciertos tipos de documentos tienen mayor número de citas que otros; este es el caso de "reviews" (re), precisamente por mostrar el estado de la cuestión de un tema en particular. Los documentos cuyas etiquetas son "lectures" (le), "notes" (no) y "conference papers" suelen tener menos referencias que los artículos propiamente dichos.

Figura 3. Diagramas de caja, correspondientes al número de citas respecto del tipo de artículo para las tres universidades gallegas

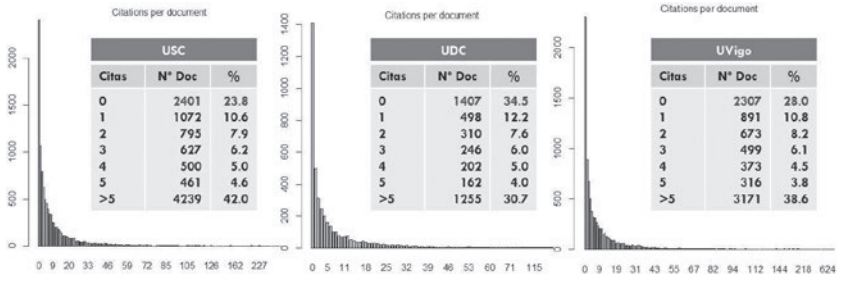


En el caso de los artículos, es paradigmático el hecho de que el número de citas a los mismos sea igual en mediana para las tres universidades (incluso en concepto de dispersión). Sin abandonar el caso de las citas a artículos,

en las tres instituciones educativas se observa una gran cantidad de este tipo de documentos, con un anormalmente alto número de citas respecto de las citas que reciben la mayoría (marcados como círculos sobre las respectivas cajas de los diagramas). Si se compara la magnitud de estos datos atípicos, según la universidad, se observa que, tanto la USC como la UVigo son capaces de producir artículos con un número de citas máximo (en torno a quinientas, o más, en el caso de la UVigo) mayor que la UDC. Esto se traduce en la existencia de una élite de investigadores en la USC y UVigo, capaces de generar artículos que despiertan más interés, ya sea por su calidad o debido a la alta producción en el campo de la ciencia donde se publican.

Si se examina la distribución de citas globalmente para cada universidad, sin atender al tipo de documento, se obtienen los resultados descritos en la *Figura 4*, es decir, la distribución del número de artículos, según el número de citas recibidas. Se observa que la USC genera, en número y proporción, más artículos con un alto número de citas que las otras dos universidades; de hecho, el 42% de los artículos producidos por la USC entre el 2000 y 2011 habían recibido más de cinco citas; mientras que sólo el 30.7y el 38.6% de los artículos en la UDC y UVigo, respectivamente, alcanzan ese impacto. Asimismo, se observa una relación inversa para la proporción de artículos con 0 citas: 23.8% para la USC, mientras que este valor es ligeramente mayor para la UVigo, 28 y, sobre todo, la UDC, 34.5%.

Figura 4. Diagramas de barras, correspondientes al número de documentos respecto del número de citas recibidas



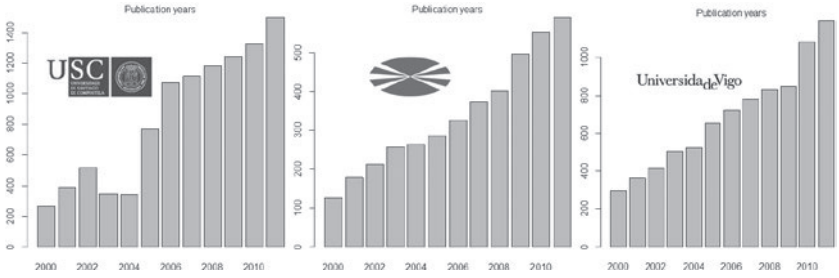
Adicionalmente, dado que el paquete Citan proporciona los agregados totales del número de citas y el número de documentos, se calcularía la proporción de citas por documento de cada universidad, índice que brinda información acerca del impacto de la investigación de una universidad en la comunidad científica y tecnológica. Así, se calcularía este índice de impacto para la USC, 101628 / 10095 = 10.1 citas/doc, la UDC, 26015 / 4080 = 6.4

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

citas/doc, y la UVigo, 78253 / 8230 = 9.5 citas/doc. Observando estos tres resultados, en términos globales y en el periodo examinado, se infiere que el impacto de los documentos generados por la USC y la UVigo es más o menos el mismo; mientras que la producción de la UDC se sitúa ligeramente por detrás. El hecho de que la distancia entre la USC y la UVigo sea menor respecto de este último índice, se debería quizás a que la UVigo ha producido en este intervalo un considerable número de artículos con un muy alto número de citas, en comparación con la USC.

En un estudio bibliométrico es de suma importancia el cálculo de la producción científica en la escala temporal o, lo que es lo mismo, la distribución del número de documentos, según el año en que se produjeron. La *Figura 5* representa el cálculo de la distribución ya mencionada para cada una de las universidades durante el periodo 2000-2011. En los tres gráficos se observa un ligero estancamiento en el crecimiento de la producción científica en la horquilla comprendida entre el 2003 y 2004.

Figura 5. Diagramas de barras, correspondientes al número de documentos respecto del año de producción

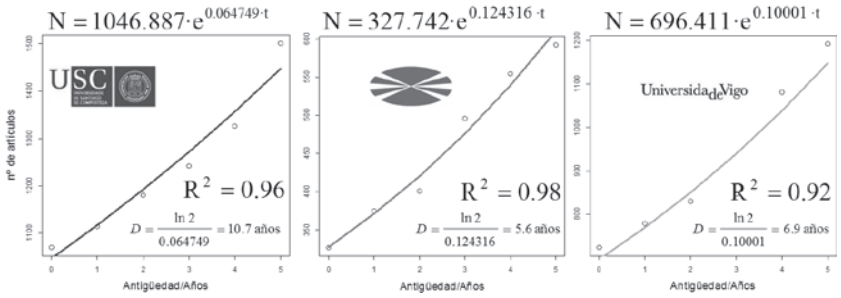


Las razones de esta tendencia quizá obedecen a recortes en número y cantidad financiada de proyectos de investigación sufragados por el Estado, pero también a una disminución del número de fuentes Scopus, pues la permanencia de algunas revistas no es continua, sino que varía a lo largo del tiempo, atendiendo a las políticas de indización. También en las tres universidades se observa un crecimiento exponencial del número de documentos en el intervalo 2006-2011, siendo el aumento en los últimos años especialmente fuerte para la UVigo y la UDC.

Llegado este punto, es necesario ir más allá del examen descriptivo y modelizar la relación entre la producción y el tiempo. En particular se aplica el modelo de crecimiento exponencial de la ciencia o Ley de Price a los datos de cada una de las universidades por separado y en el intervalo 2005-2011, después de la primera saturación de carácter sigmoide.

Se procede al ajuste de regresión no lineal a los datos y se obtienen coeficientes de determinación superiores o iguales a $R^2 = 0.92$ en todos los casos. Dado que son una medida fiable de la bondad de ajuste, se comprueba que se cumple la Ley de Price de crecimiento exponencial (*Figura 6*). Los parámetros resultantes de los ajustes del modelo de Price indican que la UDC y la UVigo son instituciones nuevas, con mayor capacidad potencial de crecimiento que la USC, de hecho, el tiempo de duplicación de la producción científica, D, es ostensiblemente menor (casi la mitad) para las dos primeras respecto de esta última (*Figura 6*).

Figura 6. Ajuste del modelo de crecimiento exponencial de Price. Cálculo del tiempo de duplicación, D

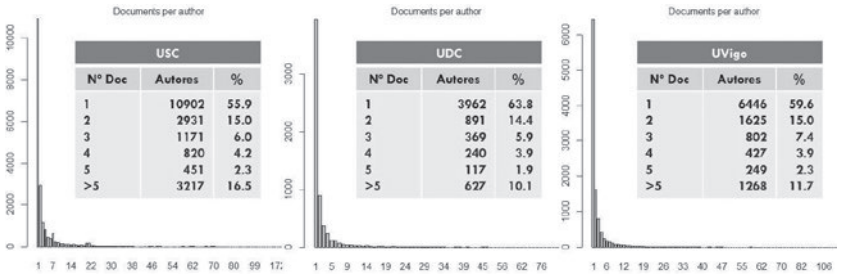


La distribución del número de autores, según la cantidad de documentos publicados, es similar en las tres universidades. En lo tocante a la USC, la proporción de autores con más de cinco artículos es ligeramente mayor que en las demás universidades; mientras que la de autores con sólo un artículo es ligeramente menor. Este resultado evidencia que la USC es una universidad más antigua, donde sus investigadores tienen una mayor trayectoria investigadora y, muy probablemente, redes de investigación más consolidadas.

La ley bibliométrica fundamental o modelo de regresión no lineal que relaciona el número de autores con el de documentos que producen es la famosa Ley de Lotka. Por tanto, el siguiente paso de este trabajo es la verificación de dicha ley respecto de los datos experimentales, ajustando el modelo no lineal mediante un algoritmo de optimización global como, por ejemplo, el algoritmo del Nelder Mead (utilizando la función de R optim), o uno de tipo evolutivo, como el cada vez más utilizado Differential Evolution (mediante la función DEoptim, dentro del paquete homónimo).

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

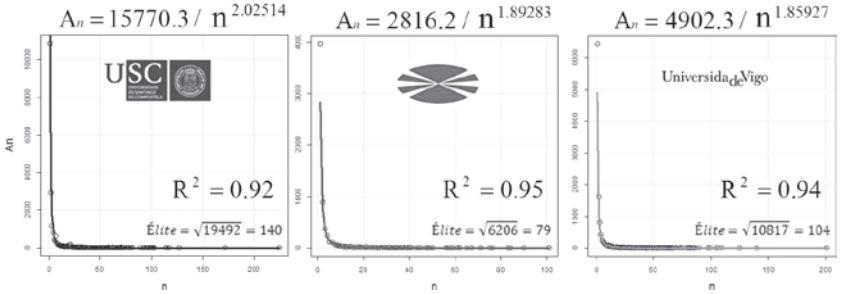
Figura 7. Diagramas de barras, correspondientes al número de autores respecto del número de documentos producidos por aquéllos



En todos los casos, se han obtenido coeficientes de determinación superiores o iguales a $R^2 = 0.92$, o lo que es lo mismo, mediante el modelo se explica al menos el 92% de la variabilidad total de los datos. Además, se obtiene que el parámetro m, correspondiente al exponente de la ley, es muy próximo al 2 teórico característico de su expresión tradicional. Por lo tanto, se comprueba el cumplimiento la Ley de Lotka en el intervalo 2000-2011.

En la *Figura 8* se observan los resultados obtenidos del ajuste de los datos. Así, el número de autores de la UVigo y de la UDC decrecen más lentamente respecto del número de documentos publicados que el número de autores en la USC. Ello significa que tiende a haber una mayor igualdad u homogeneidad en lo que respecta a la producción de los autores de las universidades más jóvenes, UDC y UVigo.

Figura 8. Resultados de la aplicación de la Ley de Lotka a los datos correspondientes a cada universidad



Para completar la información proporcionada por la Ley de Lotka, es usual aportar los resultados obtenidos de la estimación del tamaño o número de autores que forman la élite de la universidad y su red de investigación. Se

ha estimado mediante el cálculo de la raíz cuadrada del número total de autores, resultando la siguiente secuencia (siempre en torno a un valor de 100): tamaño élite (USC) > tamaño élite (UVigo) > tamaño élite (UDC). Cabe destacar que, aunque la UVigo y la UDC presentan un tamaño similar en cuanto al número de profesores, la red de investigación es mayor en la UVigo (investigadores y colegios invisibles) y, por ende, también el número de investigadores que forman parte de su élite.

En el *Cuadro 1* se muestran los diferentes índices bibliométricos calculados para cada una de las universidades. Conviene destacar que los índices elegidos, usualmente aplicados a los datos de producción científica de investigadores de manera individual, en la metodología que en este trabajo se propone se estiman para el conjunto de cada universidad.

Así, se aprovechan todas las ventajas del uso de esos índices para la evaluación de la calidad investigadora y, en consecuencia, para la comparación, no sólo de investigadores, sino de grupos, laboratorios, universidades o entidades mayores. En el *Cuadro 1* se observa que, si bien tienen similar orden de magnitud en los tres centros, la USC presenta por lo general índices más altos que las demás instituciones, ya sea que se calcule el número de citas totales, el número de documentos producidos, los índices h y g, etc.

Cuadro 1. Indicadores bibliométricos obtenidos para cada universidad dentro del periodo 2000-2011

Índice	USC	UDC	UVigo
Nº de documentos	10,095	4,080	8,230
Max nº citas	653	360	1,246
Suma de citas	101,628	26,015	78,253
h	93	49	86
g	136	71	145
rp_1	93	49	86
I_1	332.6	175.5	282.4
I_{inf}	172.8	92.09	147.5
Ge_1	7,694	2,673	5,923
Ge_5	1,997	619	1,496
SL_{p1}	16,057.1	5,050.3	12,067.1
h_h = N autores con h ≥ N	21	15	19

Los resultados indican que la USC tiene una mayor producción científica que las demás universidades y un mayor impacto en la comunidad científica. La USC es una universidad más consolidada, antigua, más asentada y conocida que la UDC o la UVigo. Sin embargo, tanto el número máximo de citas recibidas por un documento, como el índice g, han resultado mayores en la UVigo. Esto último debido al gran impacto (representado por un elevado número de citas) que presentan ciertos artículos de la UVigo.

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

Como se verá más adelante, la existencia de estos artículos se relaciona con el trabajo de investigadores de especial relevancia en campos de la ciencia muy populares en los últimos años, como la nanotecnología. Como comentario final, es especialmente relevante el hecho de que los índices calculados para la USC son muy semejantes a los de la UVigo, sobre todo si comparamos estas diferencias con las existentes respecto de la UDC, situada en tercer lugar, según todos los índices.

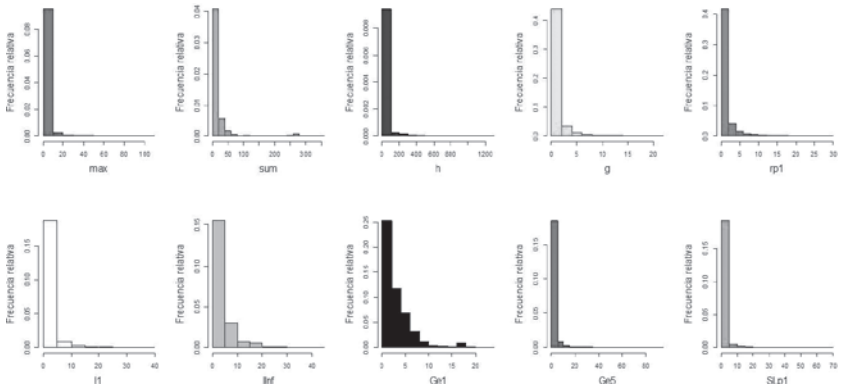
Es especialmente reseñable el cálculo del último índice de la tabla, h_b = el máximo número *N* de autores con al menos un número *h* igual a *N*. El índice bibliométrico definido como h_b es una variante del índice *h* que se estima y presenta por primera vez en este trabajo, y que define el grupo de investigadores que forman la élite más productiva de cada universidad y, a su vez, mide la calidad investigadora de esas élites representativas. Su expresión matemática es la siguiente, en la cual *h* es ahora el número *h* de cada autor perteneciente a las universidades examinadas, *H* el número *h* entendido como variable aleatoria, *i* es el número de orden y *n* el número total de autores.

Los valores resultantes de este índice indican que la élite de la USC es más numerosa y de mayor calidad que la compuesta por las otras dos universidades. De hecho, un h_b (USC) = 21 significa que hay al menos 21 investigadores de la USC que poseen un índice *h* 21, quedando el límite de entrada en la élite fijado en *h* = 21. Así, tanto el límite de entrada, como el número de investigadores que lo franquean, es más alto en la USC que en las demás universidades. El límite de inclusión en la élite de la UVigo es ligeramente más bajo, h_b (UVigo) = 19, siendo igualmente menor el número de sus componentes, 19. De nuevo se observa que la UDC queda ligeramente desmarcada de las demás: h_b (UDC) = 16, es decir, hay 16 investigadores definidos por un número *h* 16. Conviene subrayar de nuevo que los datos analizados se refieren únicamente al periodo 2000-2011.

Estudio estadístico inferencial

Como a cada autor de las universidades aquí investigadas le corresponde un valor de cada tipo de índice, y dado que cada institución educativa está compuesta por un gran número de autores diferentes, los índices bibliométricos se entenderían y estudiarían como variables aleatorias, cuyos valores estén definidos por una determinada distribución de probabilidad. En la *Figura 9* se muestran los histogramas correspondientes a los valores de los distintos números índice bibliométricos, calculados a partir de los autores de la UDC.

Figura 9. Histogramas de los indicadores bibliométricos obtenidos a partir de los datos correspondientes a los investigadores de la udc durante el periodo 2000-2011



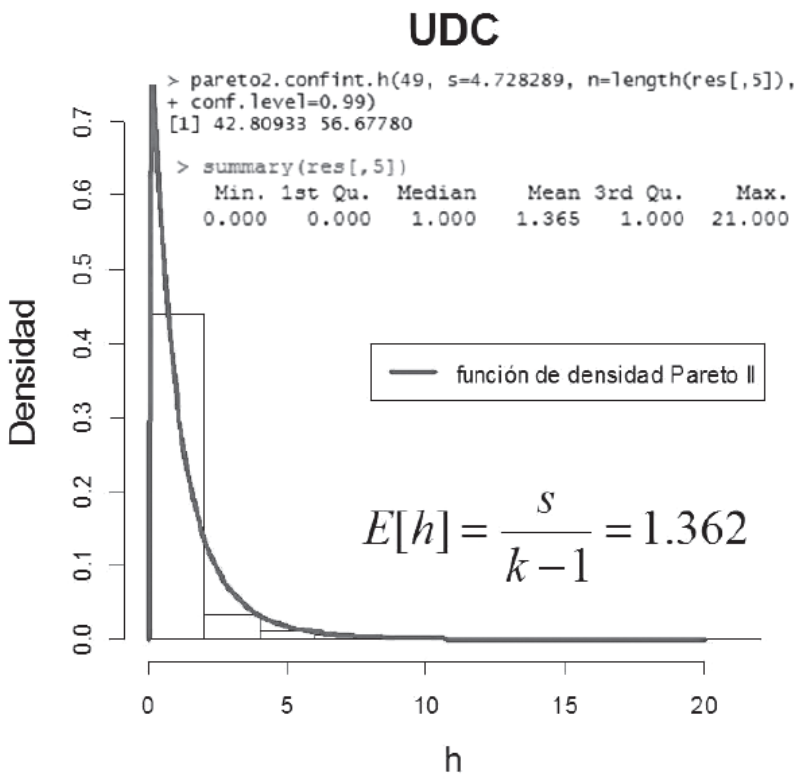
Se observa que las distribuciones de todos los índices son marcadamente asimétricas por la derecha: hay una gran cantidad de autores definidos por indicadores muy bajos, mientras que hay un número muy bajo de autores con indicadores altos (la élite de la universidad). Esta forma característica de los histogramas sugiere que la distribución de probabilidad paramétrica teórica que caracteriza los índices bibliométricos sería de tipo Pareto II.

Al igual que con el ajuste de los modelos de regresión de Price y Lotka, cuando se ajusta una distribución de probabilidad a los datos de la muestra, se pasa de la estadística descriptiva a la inferencia estadística, es decir, se hacen estimaciones de los parámetros característicos de toda la población (media, varianza, etc.), a partir de las realizaciones muestrales.

La *Figura 10* muestra el ajuste de una distribución de Pareto de tipo II a los valores del número *h* de los investigadores pertenecientes a la UDC. Los parámetros de la distribución de Pareto, tanto el de escala s como el parámetro de forma *k*, se han estimado por el método de máxima verosimilitud, mediante la aplicación de la función del paquete Citan, denominada pareto2. mleksestimate, a los datos compuestos por los números h de todos los investigadores de la UDC y de su red de investigación, denotados por el objeto res[,5].

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2017.nesp1.57891

Figura 10. Histograma de los valores del número h para los investigadores de la udc, medidas de posición muestrales, ajuste de la distribución de Pareto II, media poblacional o esperanza matemática, estimación del número h teórico de la udc mediante intervalo de confianza al 95% y aplicación del test de Wilcoxon de bondad de ajuste (código incluido) a los datos de los números h de la udc (representados por el objeto res[,5])




```
> wilcox.test(res[,5],rpareto2(length(res[,5]), k=4.471255, s=4.728289))

 Wilcoxon rank sum test with continuity correction

data:  res[, 5] and rpareto2(length(res[, 5]), k = 4.471255, s = 4.728289)
W = 18915980, p-value = 0.5047
alternative hypothesis: true location shift is not equal to 0
```

Los parámetros requeridos para la aplicación de la función son, en este orden, los datos a los que se quiere ajustar la distribución de Pareto II más próxima, la precisión deseada y un valor mínimo y máximo para el parámetro de escala: pareto2.mleksestimate(as.numeric(TotCit),tol=1e-20,smin=0.2, smax=30).

Los valores de los parámetros obtenidos por máxima verosimilitud son k = 4.471255 y s = 4.728289. Una vez estimada la distribución de Pareto II

que mejor se ajusta a los datos, convendría probar su bondad de ajuste, es decir, que estadísticamente se consideraría que la distribución de probabilidad que gobierna el valor de los datos muestrales (números *h* de los investigadores de la UDC) es una distribución de Pareto II, definida por los parámetros calculados.

La respuesta a la pregunta anterior la daría la aplicación del contraste de Wilcoxon-Mann-Whitney. El test de Wilcoxon-Mann-Whitney es una prueba no paramétrica mediante la cual se contrasta la hipótesis nula de que dos poblaciones son iguales frente a la alternativa de que son diferentes o, más comúnmente, una de estas poblaciones tiende a ser mayor que la otra.

Se obtiene que el resultado del contraste no es significativo, no se puede rechazar que los datos sigan la distribución de Pareto indicada, p-valor = 0.5047 > 0.05 (*Figura 10*), mediante la aplicación de la función wilcox.test, que compara los datos muestrales con los extraídos de la distribución de Pareto II teórica ajustada, rpareto2(length(res[,5]),k = 4.471255,s = 4.728289. Aprovechando este resultado, se haría una estimación por intervalos de confianza del verdadero número h de las universidades aquí examinadas.

Así, en el caso de la UDC, se había obtenido un valor muestral de *h* = 49; si se construye un intervalo de confianza al 95% dentro de la suposición de distribución de Pareto, el resultado es mucho más informativo, pues da una idea no sólo de posición, sino también de dispersión de la variable a estimar (*Figura 10*): (42.80933,56.67780).

Suponiendo que el índice h es en realidad una variable aleatoria, cabría cuestionarse si se consideraría que la distribución del índice h de los autores es diferente, dependiendo de la universidad o red de investigación a la que están adscritos. Para una adecuada respuesta, se aplicaría, de nuevo, el test no paramétrico de Wilcoxon-Mann-Whitney. El contraste de hipótesis definido es el siguiente:

- H_0: los índices h de los autores de las dos universidades que se comparan son iguales.
- H_1: los índices h de los autores de la primera universidad son menores que los correspondientes a los autores de la segunda.

El resultado correspondiente a la aplicación de contrastes múltiples, dos a dos, se muestra en el *Cuadro 2*. Todas las comparaciones han resultado significativas en un nivel de significación, previa corrección de Bonferroni, igual a 0.05/3 = 0.015.

DOI: http://dx.doi.org/10.22201/iibi.2448321xe.2017.nesp1.57891

Cuadro 2. Resultado del contraste de Wilcoxon aplicado para comparar las distribuciones del índice h de los autores, según su universidad de adscripción

Comparaciones	Estatístico	p-valor	Resultado
h (UDC) - h (USC)	50440118	$2.2 \cdot 10^{-16}$	Significativo
h (UDC) - h (UVigo)	30671459	$2.2 \cdot 10^{-16}$	Significativo
h (UVigo) - h (USC)	95935123	$2.2 \cdot 10^{-16}$	Significativo

Por tanto, se llega concluye que, por lo general, los autores de la UDC tienen un menor índice h que los autores de la UVigo, y los autores de la UVigo tienden a tener un índice h menor que el de los autores de la USC y su red de investigación. Con este último resultado, aparte de comparar las distintas universidades mediante el cálculo de índices bibliométricos en su conjunto, se comparan dichos centros de investigación atendiendo a los valores de los índices de cada uno de los autores que las componen. La secuencia ordenada de los números *h* para los autores de las tres universidades coincide con los resultados obtenidos en el *Cuadro 1*. La USC es una universidad más antigua, más asentada, más grande y con una mayor implantación en la comunidad científica; de ahí que sus autores tienden a tener un número *h* mayor que el de los pertenecientes a las demás universidades gallegas, hablando en términos poblacionales.

Cabe destacar también que el hecho de que los investigadores de la UVigo tiendan a tener números *h* mayores que sus colegas de la UDC es un indicador de una mayor productividad e influencia en la comunidad científica abarcada por la base de datos Scopus.

Este último dato es especialmente significativo, ya que las dos universidades poseen similar tamaño y antigüedad. Entre las posibles causas de esta situación, se podría señalar el hecho de que la UVigo está más orientada que la UDC a las ciencias experimentales e ingeniería, campos muy productivos y con numerosas fuentes incluidas en Scopus.

Conclusiones

Este trabajo presenta una nueva metodología de análisis de la información bibliométrica para evaluar centros de investigación, desde un punto de vista estadístico y utilizando el software libre R, en particular el paquete Citan. El procedimiento aquí utilizado permite un tratamiento estadístico más completo a partir de bases de datos mucho más extensas, sin limitación del número de entradas (hasta hace poco dos mil para Scopus y quinientas para el ISI). Esta característica hace que el enfoque propuesto sea particularmente

atractivo para el estudio y comparación a lo largo del tiempo de entidades con una gran producción científica, por ejemplo, las universidades.

Una parte importante de la metodología propuesta se relaciona con el análisis estadístico descriptivo de la producción científica correspondiente a las tres universidades. Como resultado, se afirmaría que la red de investigación de la USC es más productiva e influyente que la de la UDC y la UVigo, ya que casi todos sus índices bibliométricos son superiores. La USC es una universidad más antigua, conocida, con grupos de investigación más numerosos, grandes y consolidados, que a su vez se caracterizan por estar compuestos por un mayor número de autores, que además generan un mayor número de documentos, tendiendo a tener estos documentos un mayor número de citas.

Por otro lado, es especialmente representativo el caso de la UVigo que, siendo sensiblemente más pequeña, se caracteriza por índices muy próximos a los de la USC, en algunos casos mayores, como su índice g. Esto debido a que la UVigo es una universidad que conjuga una alta productividad, mayor que la de la USC, con un alto impacto en la comunidad científica.

Aparte de medir la productividad desde un punto de vista descriptivo, en este artículo se ha medido y modelado el crecimiento de la producción de las tres universidades gallegas en el periodo 2006-2011. Se ha verificado el cumplimiento de la ley de crecimiento exponencial de la ciencia, estimándose que la UDC, seguida por la UVigo, son las instituciones con una mayor velocidad de crecimiento. Particularmente, la UDC es la que presenta un mayor potencial del crecimiento de su producción, caracterizado por un tiempo de duplicación de sólo 5.6 años.

Igualmente se ha ajustado y comprobado la bondad de ajuste de la Ley de Lotka para las tres universidades. La comparación entre los parámetros m de los modelos estimados lleva a concluir que tiende a existir una mayor homogeneidad en la producción de los autores correspondientes a las universidades más jóvenes, UDC y UVigo.

Una de las principales aportaciones de este estudio es la presentación del índice bibliométrico, definido como h_b, máximo número de investigadores con un número h igual o superior a h_b, que define el grupo de investigadores que conforman la élite más productiva de cada universidad y, a su vez, mide la calidad investigadora de esas élites representativas.

Según este índice, la USC posee la élite más números y de mayor calidad, pues está compuesta por 21 autores, con un índice h al menos de 21. La UVigo se sitúa a una corta distancia con $h_b = 19$, mientras que la élite de la UDC está compuesta por 16 investigadores, con al menos un índice h de 16. Una vez más, se muestra que la UDC es la universidad con un mayor margen de mejora.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2017.nesp1.57891

Además de comparar los centros de investigación, atendiendo a los índices calculados individualmente sobre el total de su producción, también se evaluarían calculando los índices bibliométricos para cada autor y acto seguido considerar estos valores como realizaciones de variables aleatorias. La comparación de las distribuciones de los números e índices estudiados como variables aleatorias permite, a su vez, comparar las universidades.

Así, desde un punto de vista de inferencia estadística, mediante la aplicación del test no paramétrico de Wilcoxon-Mann-Whitney, se observó que los autores de la UDC tienden a tener un índice h más bajo que los correspondientes a la UVigo, mientras que los de la UVigo suelen presentar un índice h menor que los de la USC: $h(\text{USC}) > h(\text{UVigo}) > h(\text{UDC})$. Estos resultados apoyan los obtenidos por agregación para el conjunto de las universidades.

Agradecimientos

Este estudio se realizó con financiamiento aportado por los proyectos MTM2011-22392 y MTM2014-52876-R.

Referencias

Abdi, H. 2007. "The Bonferroni and Šidák Corrections for Multiple Comparisons". en *Encyclopedia of measurement and statistics*. Thounsand Oaks, Cal.:Sage.

Alcaín Partearroyo, M. D. 1991. "Aspectos métricos de la información científica", *Ciencias de la Información*, vol. 160: 32-36.

Arencibia, J. R. y R. Carvajal Espino. 2008. "Los índices H, G y R: su uso para identificar autores líderes en el área de la comunicación durante el periodo 2001-2006", *Acimed*, vol. 17, no. 4.

Arroyo, N., J.L. Ortega, V. Pareja, J.A. Prieto e I. Aguillo. 2005. "Cibermetría. Estado de la cuestión". Barcelona: Fesabid, XIX Jornadas Españolas de Documentación.

Brookes, B.C. 1990. "Biblio-, Sciento-, Infor- Metrics. What are talking about?", *Informetrics*, nos. 89-90: 31-43.

Center for Research Libraries. 2014. <http://www.crl.edu/>.

Egghe, L. 2006. "Theory and practise of the g-index", *Scientometrics*, vol. 69, no. 1: 131-152.

Fay, M. P. y M.A. Proschan. 2010. "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis test and multiple interpretations of decision rules", *Statistics Surveys*, vol. 4: 1-39.

Gagolewski, M. 2011. "Bibliometric impact assessment with R and the Citan package", *Journal of Informetrics*, vol. 5, no. 4: 678-692.

CitanGagolewski, M. y P. Grzegorzewski. 2010. "S-Statistics and their basic properties", en C. Borgelt et al., eds., *Combining soft computing and statistical methods in data analysis. Advances in Intelligent and Soft Computing*. Berlín: Springer.

Garfield, E. 1955. "Citation indexes to science: a new dimension in documentation through the association of ideas", *Science*, vol. 122: 108-111.

Hirsch, J.E. 2005. "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46: 16569-16572.

Hulme, E. W. (1923). *Statistical bibliography in relation to the growth of modern civilization*. Londres: Grafton.

Lotka, A. J. 1926. "The frequency distribution of scientific productivity", *Journal of the Washington Academy of Sciences*, vol. 16, no. 12: 317-323.

Mann, H. B. y D.R. Whitney. 1947. "On a test of whether one of two random variables is stochastically larger than the other", *Annals of Mathematical Statistics*, vol. 18, no. 1: 50-60.

Moya-Anegón, Félix de (dir.). 2013. *Indicadores bibliométricos de la actividad científica española 2010*. Madrid: FECYT, en <http://icono.fecyt.es/informesypublicaciones/Documents/indicadores%20bibliometricos_web.pdf>, consultada el 22 de abril de 2015.

Osca-Lluch, J., S. Miguel, C. González, M. Peñaranda-Ortega y E. Quiñones-Vidal. 2013. "Cobertura y solapamiento de Web of Science y Scopus en el análisis de la actividad científica española en psicología", *Anales de Psicología*, vol. 29, 3: 1025-1031.

Price, D. J. 1973. *Hacia una ciencia de la ciencia.* Barcelona: Ariel [1963].

Priem, J., D. Taraborelli, P. Groth y C. Neylon. 2010. "Altmetrics: a manifesto – altmetrics.org", en <http://altmetrics.org/manifesto/>, consultada el 15 septiembre de 2014.

Pritchard, A. 1969. "Statistical bibliography; an interim bibliography", North-Western Polytechnic, School of Librarianship, vol. 60, no. 5: 184-244.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, en <http://www.R-project.org>, consultada el 20 abril de 2015.

Spinak, E. 1996. *Diccionario enciclopédico de Bibliometría, Cienciometría e Informetría*. Caracas: Unesco.

Storn, R.M., K. Price y J.A. Lampinen. 2005. *Differential evolution: A practical approach to global optimization*. Berlín: Springer.

Para citar este texto:

Tarrío-Saavedra, Javier; Orois, Elena; Naya, Salvador. 2017. "Estudio métrico sobre la actividad investigadora usando el software libre R: el caso del sistema universitario gallego". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información* (Número Especial de Bibliometría): 221-247.

http://dx.doi.org/10.22201/iibi.24488321xe.2017.nesp1.57891

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2017.nesp1.57891