Revista Española de Documentación Científica
38(4), octubre-diciembre 2015, e102
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2015.4.1225

ESTUDIOS / *RESEARCH STUDIES*

Diferencias y evolución del impacto académico en los perfiles de Google Scholar Citations: Una aplicación de árboles de decisión

José Luis Ortega*

* Laboratorio de Cibermetría, CCHS-CSIC, Madrid, España
Correo-e: jortega@orgc.csic.es

Recibido: 19-09-2014; 2ª versión: 24-03-2015; Aceptado: 10-04-2015.

Cómo citar este artículo/Citation: Ortega, J. L. (2015). Diferencias y evolución del impacto académico en los perfiles de Google Scholar Citations: Una aplicación de árboles de decisión. *Revista Española de Documentación Científica*, 38(4): e102. doi: http://dx.doi.org/10.3989/redc.2015.4.1225

Resumen: El propósito de este artículo es analizar la producción e impacto de más de 3000 perfiles tomados de Google Scholar Citations con el fin de identificar qué segmentos (por género, puestos académicos y disciplinas) son más exitosos en términos de impacto científico. Este análisis se afrontó tanto desde una perspectiva estática como longitudinal. Los árboles de decisión fueron usados para detectar las variables más importantes para agrupar perfiles con un mayor número de citas por artículo e índice h. Resultados muestran que la carrera académica es el factor más importante para conseguir citas y mejorar el índice h. Los investigadores más veteranos son así los que ocupan las primeras posiciones, mientras que los jóvenes investigadores describen curriculums en ciernes. Por el contrario, estos resultados cambian cuando el crecimiento de los perfiles es observado. Así los curriculums más jóvenes son los que experimentan un crecimiento más fuerte, mientras que los más veteranos muestran signos de estabilización y estancamiento. Se concluye que los investigadores con una carrera estable pertenecientes a las ciencias de la vida tienen mejor impacto que los jóvenes investigadores de humanidades y ciencias sociales, a pesar de que estos últimos son los que más rápido crecen en número de citas por documento.

Palabras clave: Bibliometría; Google Scholar Citations; buscadores académicos; árboles de decisión; impacto científico.

Differences and evolution of scholarly impact in Google Scholar Citations profiles: An application of decision trees

Abstract: The aim of this paper is to analyse the research performance of more than 3,000 profiles from Google Scholar Citations to define which groups (by gender, academic positions and disciplines) bring together more successful profiles. This analysis was faced both from a static and a longitudinal point of view. Decision trees were used to detect the most important variables in order to distinguish winning profiles and to observe which categories bring together more authors with high number of citations and h-indexes. Results show that the career is the most relevant aspect to achieve citations and improve the h-index. Senior researchers are thus ranked in the best positions, while young scholars describe nascent curricula. Otherwise, this distribution changes when growth rates are computed. It is concluded that researchers with a stable career from life sciences have better research impact than young researchers from humanities and social sciences, despite that the fastest growing profiles belong to young scholars.

Keywords: Bibliometrics; Google Scholar Citations; academic search engines; decision trees, research impact.

1. INTRODUCCIÓN

La evaluación de la ciencia, basada en indicadores bibliométricos, se ha topado con el hecho de que la actividad científica no es uniforme; al contrario, existen diferentes maneras de llevar a cabo una investigación y cuyos resultados pueden ser transmitidos de diferentes formas. La ciencia básica está orientada a la publicación de resultados teóricos (i.e. artículos, libros, etc.) dirigidos a una audiencia especializada, mientras que las ciencias aplicadas desarrollan instrumentos y aplicaciones (i.e. patentes, modelos de utilidad, etc.) con un propósito económico o social (Gibbons y otros, 1994). Además, los investigadores también están inmersos en actividades de formación y divulgación que complementan su actividad científica, produciendo resultados orientados a un público más general (Kidd, 1988).

De este modo, esta gran variedad de resultados académicos también provocan que las citas provengan de diferentes entornos y sean generadas por múltiple razones. Así el impacto científico de un investigador depende de la línea de investigación que desarrolle, el tiempo que esté implicado en esos estudios, o el país donde desarrolle su actividad. Esto provoca serios problemas en la comparación de distintas trayectorias científicas provenientes de distintos entornos académicos, además de conocer qué condiciones influyen en el éxito de un investigador en términos bibliométricos. Más aún, es difícil entender cómo estos aspectos pueden favorecer el desarrollo de una carrera prometedora. A través del uso de árboles de decisión, este trabajo intenta explorar qué atributos caracterizan el impacto de un científico y por lo tanto diferenciar perfiles en procesos de evaluación.

Por otro lado, el impacto que estos productos tienen tanto en la comunidad científica como en la sociedad en general es muy diferente y, sobretodo, difícil de medir. Debido a esto, la bibliometría se está abriendo a nuevos indicadores (i.e. altmetricos, cibermetricos, etc.) y a nuevas fuentes de datos (i.e. redes sociales académicas, buscadores académicos, etc.) que amplíen el ámbito del impacto científico (Aguillo y otros, 2005; Piwowar, 2013). En este contexto, este estudio introduce el uso de Google Scholar Citations (GSC) como una nueva fuente abierta para explorar su idoneidad para análisis bibliométricos y su adaptación a la evaluación científica.

2. ESTUDIOS RELACIONADOS

Unos de los propósitos iniciales de la bibliometría ha sido descubrir qué elementos influyen en la obtención de citas. Desde un primer momento ya se detectaron diferencias significativas en la distribución de citas por disciplinas, como causa de diferentes culturas científicas (Solla Price, 1970; Small y Griffith, 1974). Esto animó a diferentes estudios a explorar la razones detrás de estas diferencias y sus implicaciones para la evaluación científica (Kostoff, 1998). En este sentido, muchos artículos argumentaron que estas diferencias eran debidas a efectos de tamaño tal como número de publicaciones (Small y Crane, 1979; Schubert y Braun, 1986), referencias (Garfield, 1980) o grado de colaboración (Smart y Bayer, 1986). Recientemente, estos patrones de citación en disciplinas han sido estudiados desde un enfoque evolutivo. Radicchi y otros (2008) observaron que la evolución de las citas al nivel de artículo describe una distribución de carácter general siempre cuando estas son normalizadas por el promedio de citas en la disciplina. Althouse y otros (2009) analizaron la razón del incremento del factor de impacto en revistas y detectaron que esta se debía, en gran medida, a cambios en la longitud de la lista de referencias. Finalmente, Finardi (2014) apreció diferentes patrones de crecimiento en diferentes revistas desde la química a las ciencias sociales.

Pero, quizás, estas diferencias se aprecian mejor cuando el puesto académico de un investigador es considerado como un indicador de la madurez investigadora. Muchos trabajos han tratado estas diferencias principalmente con respecto a la producción científica (Long, 1978; Hancock y otros, 1992; Jacobs y Ingwersen, 2000); mientras que sólo unos cuantos han apuntado a la relación entre carrera y citaciones. Ventura y Mombrú (2006) testaron la actividad de Profesores y Profesores Asociados en Uruguay y detectaron que el puesto académico afectaba positivamente sus ratios de citas. Abramo y otros (2009), estudiando a 33.000 investigadores italianos, observaron que el impacto de estos se incrementaba a medida que conseguían mejores puestos académicos. Pagel y Hudetz (2011) obtuvieron resultados similares cuando analizaron el índice h de más de 1.600 anestesiólogos norteamericanos. Sin embargo, Aksnes y otros (2011) estudiaron 8.500 investigadores noruegos para extrañamente no encontrar diferencias significativas entre puestos académicos. Desde una óptica evolutiva, Penner y otros (2013) concluyeron que la actividad científica dependía enormemente del tiempo que los investigadores pasaban ocupando un puesto académico y que los primeros años eran cruciales para construir un futuro curriculum exitoso (Maranto y Streuly, 1994).

Por otro lado, la literatura sobre diferencias de género ha sido más prolífica. Muchos estudios han

detectado variaciones en el número de artículos científicos (Kyvik y Teigen, 1996; Abramo y otros, 2009) y puestos académicos (Long, 2001) por parte de hombres y mujeres. Aunque estas diferencias no se observaron a la hora de conseguir citas (Ding y otros, 2006; Penas y Willett, 2006). Aksnes y otros (2011) sí detectaron diferencias de género pero esta vez fueron motivadas por factores de producción y ventajas acumulativas.

El lanzamiento de GSC en 2011 atrajo la atención de distintos investigadores con el fin de explorar el potencial de esta herramienta para la evaluación científica (Pitney y Gilson, 2012; Huang y Yuan, 2012). Ortega y Aguillo (2012) construyeron un Mapa de la Ciencia usando las etiquetas incluidas en cada perfil. Ellos también mapearon la red de colaboración entre países e instituciones usando la lista de co-autores incluida en cada perfil (Ortega y Aguillo, 2013). Por otro lado, Delgado López-Cózar y otros (2014) evidenciaron la posibilidad de manipular las puntuaciones bibliométricas dentro de los perfiles. Con respecto a Google Scholar, muchos trabajos han estudiado su cobertura en relación a otras bases de datos (Meho y Yang, 2007; Kousha y Thelwall, 2007). Chen (2010) describió su rápido crecimiento en relación a diversas bases de datos. Más recientemente, Orduña y Delgado López-Cózar (2014) detectaron que en menos de un año el índice h de todas las revistas en Google Scholar aumentó un 15%.

3. OBJETIVOS

El propósito de este estudio es analizar la evolución de las citas recibidas por parte de más de 3.000 perfiles de GSC a partir de distintas muestras extraídas durante el periodo 2011-2013. Se pretende de esta forma describir qué factores (género, puesto y área de investigación) podrían influenciar en la evolución de estos indicadores bibliométricos. Este objetivo principal es detallado a través de diferentes preguntas:

- ¿Existe alguna diferencia de género, puesto o entre áreas de investigación a la hora de conseguir un mayor impacto científico?
- ¿Existe alguna diferencia de género, puesto o entre áreas de investigación que afecten en mayor o menor medida la evolución de este impacto científico?
- ¿Son los árboles de decisión herramientas adecuadas para distinguir y clasificar la actividad investigadora de los autores?
- ¿Es Google Scholar Citations un instrumento apropiado para el análisis bibliométricos?

4. MÉTODO

Extracción de datos

Google Scholar Citations es un servicio web que facilita la publicación de curricula personales a partir de datos bibliográficos tomados de Google Scholar. Junto a esta lista de publicaciones, el servicio calcula distintos indicadores bibliometricos (citas, índice h, etc.) y muestra datos de identificación (nombre, afiliación, dominio de correo electrónico, etc.). Este servicio fue lanzado en noviembre de 2011 y probablemente contenga cerca de 300.000 perfiles en diciembre de 2013 (Ortega, en prensa). Las razones para seleccionar esta fuente de datos fueron:

- Es un servicio abierto y gratuito que permite la extracción automática de datos de estos perfiles.
- Su rápida actualización favorece la extracción de distintas muestras a lo largo del tiempo y su comparación.
- En algunos casos, es posible identificar el puesto, género y área de investigación de cada perfil lo que facilita la agrupación de perfiles por categorías.
- Google Scholar es probablemente la base de datos científica más exhaustiva por lo que sus resultados pueden ser más consistentes y fiables.

En trabajos anteriores se detalló el proceso de recolección de datos (Ortega y Aguillo, 2012; 2013). Aún así, este se desarrolló en dos fases: en la primera se escribió un script en SQL para rastrear el sitio preguntando por las 26 letras del alfabeto latino en grupo de dos, identificando a tantos perfiles como fueran posible y extrayendo sus códigos de identificación. Una vez que este proceso de identificación fue terminado, un segundo script recababa los datos identificativos de cada perfil, como son nombre, afiliación, etc.; e indicadores bibliométricos (citas, documentos, índice h e índice i10). Cinco muestras trimestrales fueron obtenidas desde diciembre de 2011 a diciembre de 2012, junto a otra anual en diciembre 2013.

12.480 perfiles presentes en todas las muestras fueron tomados para comprobar la evolución bibliométrica de estos investigadores. A continuación, estos registros fueron enviados a un proceso de limpieza y normalización para homogeneizar y agrupar las variables categóricas. A partir de esta fase, sólo 3.034 perfiles pudieron ser clasificados simultáneamente de acuerdo a las siguientes categorías:

- Género: Se identificó el género de 7.673 (61%) perfiles utilizando su primer nombre. Esto fue posible para los nombres más frecuentes y usuales para hombre y mujer. En caso de duda no se asignó género.
- Puesto: seis categorías profesionales, tan próximas como fuese posible a la jerarquía académica, se definieron para agrupar las escalas académicas. Este apartado fue rellenado sólo en el caso donde existía una mención explícita al puesto de trabajo de cada perfil, por ejemplo, "Candidate Ph.D of Computer Science, QUT". De este modo se identificaron 6.559 (52,5%) puestos:
  - Becarios pre-doctorales (*Doctoral Students*): estudiantes pre-doctorales
  - Ayudantes de Investigación (*Research fellow*): personal técnico de apoyo a la investigación e investigadores contratados
  - Profesores Asistentes (*Assistant professors*): ayudantes de profesores
  - Profesores Asociados (*Associate professors*): profesores y científicos contratados que no poseen la titularidad
  - Profesores (*Professors*): profesores titulares y científicos de plantilla
  - Profesores Eméritos (*Emeritus professors*)
- Área de investigación: como en previos estudios sobre GSC (Ortega y Aguillo, 2012), las etiquetas de cada perfil fueron agrupadas y clasificadas para describir el interés científico de cada investigador. Subject Area categories de Scopus (2014) fue usado para agrupar las etiquetas en cuatro áreas científicas principales: Ciencias Físicas, Ciencias de la Salud, Ciencias Sociales y Ciencias de la Vida. Artes y Humanidades fue añadida porque se supuso que estos investigadores podrían presentar un comportamiento diferente a los de Ciencias Sociales y por lo tanto ambas deberían ser analizadas por separado. 8.743 (70%) perfiles pudieron ser clasificados.

Indicadores

A continuación se calcularon dos indicadores bibliométricos que expresasen un valor relativo entre producción (artículos) e impacto (citas). Estos indicadores fueron considerados más robustos porque fueron construidos como ratio entre dos magnitudes interrelacionadas y dependientes:

- Cit./Art.: Cantidad total de citas recibidas por cada autor dividido por el número de artículos.
- Índice h: Se define formalmente como el número de artículos *h* que han recibido como mínimo *h* citas. Por ejemplo, un índice h=5 significa que un autor ha publicado al menos 5 artículos que han sido citados cinco o más veces. Sin embargo, este indicador es muy dependiente del número de publicaciones.

Se calculó una medida de crecimiento (C_t) que cuantifique el incremento trimestral de estos indicadores desde diciembre de 2011 a diciembre de 2013. La fórmula del interés compuesto fue usada para describir el crecimiento promedio de los indicadores bibliométricos de un modo porcentual.

$$C_t = \left[\left(\frac{V_1}{V_n} \right)^{\frac{1}{n}} - 1 \right] * 100$$

Donde V_1 es el valor inicial, V_n el valor final y n es el número de momentos que van desde la observación inicial a la final.

Árboles de decisión

Esta es una técnica estadística ampliamente usada en minería de datos que permite agrupar elementos descritos por una variable (dependiente) en función de los valores de otras variables independientes (predictores). Su objetivo es trazar variaciones significativas en la distribución de la variable dependiente con respecto a las otras variables independientes, caracterizando qué factores tienen más influencia en la detección de grupos homogéneos. Este proceso es desarrollado a través de un proceso de razonamiento en el que un algoritmo (CHAID, CRT, QUEST, etc.) detecta la variable más influyente sobre la variable dependiente, dividiendo el original nodo y construyendo un árbol de nuevos nodos que a su vez clasifican las observaciones con respecto a la variable dependiente. Este proceso continúa hasta que cada grupo describe la máxima pureza, esto es, cada grupo contiene sólo la mayor proporción de un único valor de la variable dependiente. De este modo, es posible conocer qué valores de una variable afectan significativamente a la distribución de la variable dependiente, construyendo perfiles de objetos o personas. Esta técnica es adecuada para variables nominales u ordinales porque es más fácil de observar cómo la presencia y ausencia del valor de una variable puede afectar a la distribución de la muestra. El algoritmo exhaustivo CHAID (Chi-square automatic interaction detector) se usó debido a que es el

más generalizado y restrictivo en sus resultados. Este algoritmo utiliza el test de chi cuadrado para generar nuevos nodos a partir de las diferencias detectadas en la distribución.

Las variables bibliométricas (Cit./Art. e índice h) se transformaron de continuas a ordinales para implementar esta técnica y obtener una mejor interpretación de los resultados. Estas variables fueron de este modo ordenadas y agrupadas en cuartiles. Así, el cuartil 1 correspondería con el 25% de los perfiles con puntuaciones bibliométrcicas más altas, mientras que el cuartil 4 agruparía al 25% con peores resultados.

5. RESULTADOS

Los árboles de decisión se usaron para encontrar qué tipo de perfiles consiguen mejores resultados en términos bibliométricos. Sin embargo, los resultados describen grupos con baja pureza porque las variables objeto (cuartiles de citas e índice h) no son enteramente categóricas por lo que sus valores no son exclusivos. Esto quiere decir que cualquier grupo (hombres, mujeres, profesores, ciencias sociales, etc.) puede tener miembros en los cuatro cuartiles. Así pues el objetivo de esta técnica en este estudio es sólo observar visualmente cómo el impacto científico se distribuye en función de puestos, género y áreas científicas, y no simplemente construir un modelo de clasificación con alta pureza y bajo riesgo. Debido a esto, los valores de riesgo son generalmente altos (riesgo>.5) y los grupos tienden a estar equilibrados. A pesar de esto, se consideró un p-valor>.005 para determinar cada rama con una aceptable significación estadística.

Actividad actual

En esta sección, se observa la actividad actual para luego ser comparada entre la actividad acumulada de un investigador y cómo esta evoluciona durante tres años. El momento de referencia usado es el más actualizado, diciembre de 2013. La Tabla I muestra los intervalos y número de casos por cuartil, lo que facilita la interpretación de los resultados.

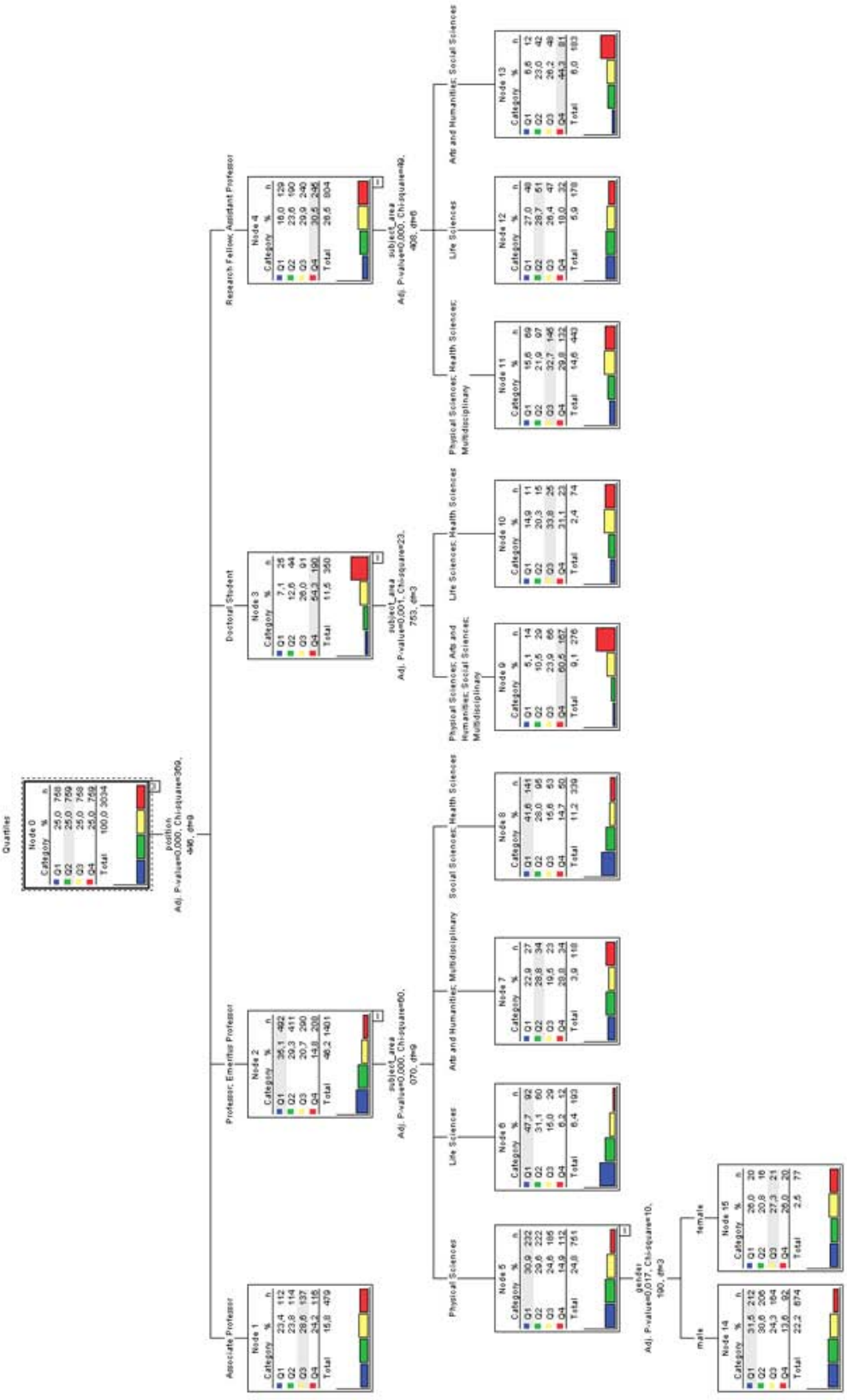
Figura 1 muestra el árbol de decisión para los cuartiles del número de citas por documento (Cit./Art.). La variable que más influencia la distribución de cuartiles es el puesto académico. Así, el 35% de los *Profesores-Profesores Eméritos* están dispuestos en el primer cuartil; mientras que el 54% de los *Becarios pre-doctorales* están localizados en el cuarto cuartil, mostrando así el resultado académico más bajo. Descendiendo a la siguiente rama, la segunda variable en importancia es área de investigación. Según esta, los autores más exitosos son *Profesores-Profesores Eméritos* de *Ciencias de la Vida* porque ellos contienen el 47.7% de los investigadores en Q1, seguido por *Ciencias Sociales-Ciencias de la Salud* con el 41.6%. Por otro lado, autores con los resultados más bajos son *Becarios pre-doctorales* de *Ciencias Físicas-Arte y Humanidades-Ciencias Sociales-Multidisciplinar*, con el 60.5% de los casos en Q4, junto a *Ayudantes de Investigación-Profesores Asistentes* de *Artes y Humanidades-Ciencias Sociales* con el 44.3% también en Q4. La tercera variable, género, es poco relevante y sólo detecta diferencias en el caso de *Profesores-Profesores Eméritos* de *Ciencias Físicas*. En este caso, el 31.5% de los hombres están en el Q1 y el 13.6% en Q4, mientas que el 26% de las mujeres están tanto en Q1 como en Q4.

De acuerdo al índice h, el árbol de decisión aporta una configuración considerablemente diferente con diferencias más destacables (Figura 2). Los puestos académicos siguen siendo aún la variable más importante, donde el 44.8% de los *Profesores-Profesores Eméritos* están localizados en el primer cuartil, y el 84.3% de los *Becarios pre-doctorales* están en el Q4 con un índice h por debajo de 7. La segunda rama está, en algunos casos, formada por criterios de género y disciplinas. Esto puede ser debido a la baja presencia de mujeres en la muestra, lo que causaría que no tuvieran el suficiente poder discriminador en algunos grupos. Esta baja presencia de mujeres en la muestra explicaría que en la rama de *Profesores-Profesores Eméritos*, los hombres tengan casi el doble de autores en el cuartil primero (47.1%) que las mujeres (26.3%). En función

Tabla I. Distribución de Cit./Art. e índice h por cuartiles en la actividad actual

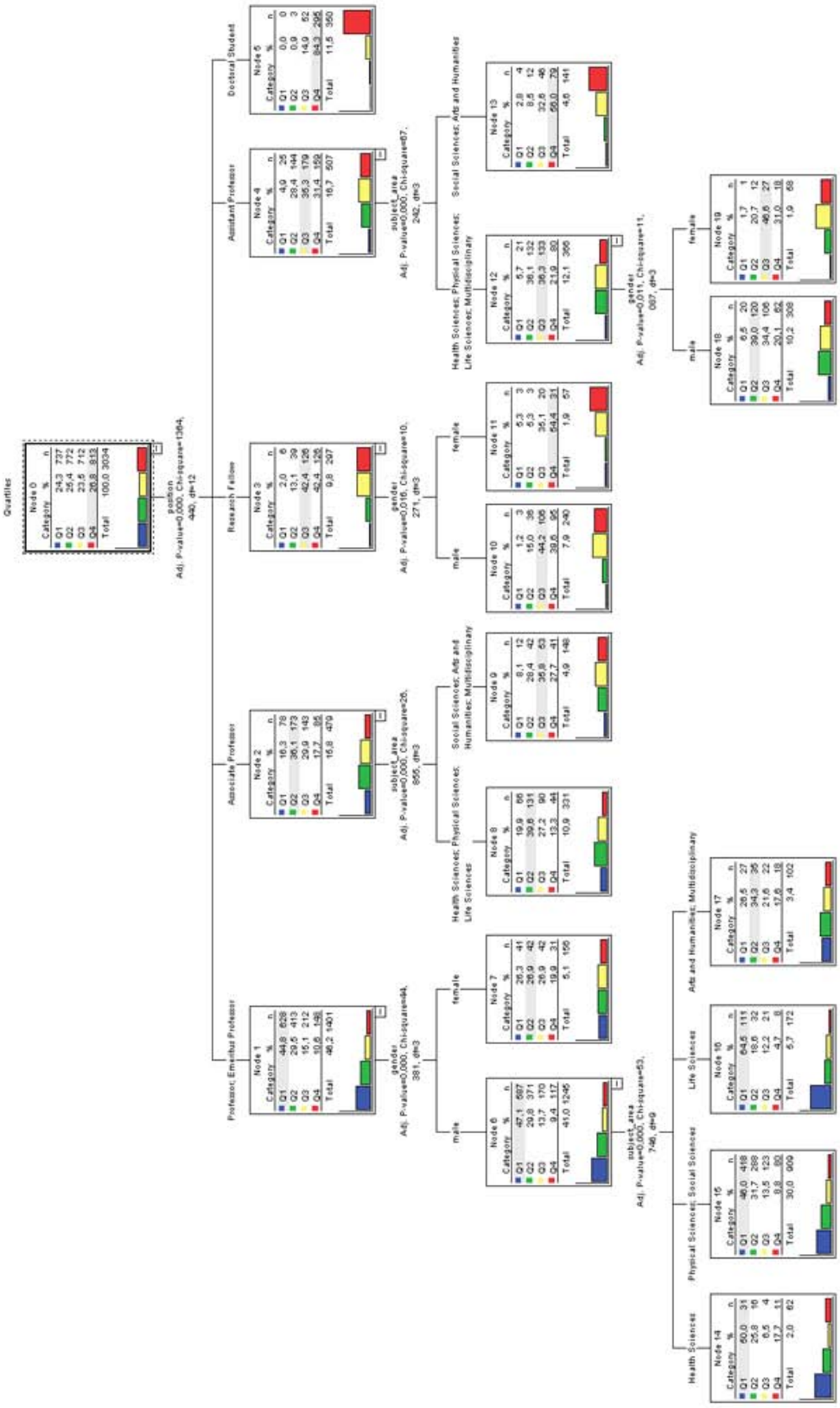
Cuartil	*Cit./Art.*	*N*	*Índice h*	*N*
Q1	30.3 - 891.18	758	28 - 234	737
Q2	14.24 - 30.2	759	15 - 27	772
Q3	6.78 - 14.22	758	8 – 14	712
Q4	0 - 6.77	759	0 – 7	813
Total		3034		3034

Figura 1. Árbol de decisión de acuerdo a cuartiles de Cit./Art.



Rev. Esp. Doc. Cient., 38(4), octubre-diciembre 2015, e102. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2015.4.1225

Figura 2. Árbol de decisión según cuartiles de índice h



de la áreas científicas, los investigadores más exitosos son *Profesores-Profesores Eméritos*, hombres de *Ciencias de la Vida* (64.5%), *Ciencias de la Salud* (50%) y *Ciencias Físicas-Ciencias Sociales* (46%). Por otro lado, los investigadores con los ratios más bajos de índice h (Q4) son *Ayudantes de Investigación*, mujeres (54.4%) y *Profesores Asistentes* de *Ciencias Sociales-Arte y Humanidades* (56%).

Evolución de la actividad

Después de describir el perfil actual de los investigadores en GSC, esta sección presenta cómo estos perfiles han evolucionado en función de su género, puesto y área de investigación. Los incrementos son medidos como promedios trimestrales durante dos años. La Tabla II describe los intervalos y número de casos por cuartil. El crecimiento es medido en porcentajes.

Figura 3 despliega el árbol de decisión según el grado de crecimiento del ratio citas por artículo. Como en los árboles anteriores, el puesto académico es el principal criterio para desplegar las ramas del árbol. Pero de forma contraria a como lo hacía en Actividad actual, *Becarios pre-doctorales* (59.7%) y *Ayudantes de Investigación* (49.5%) son los investigadores que más incrementan su ratio de citas por artículo; mientras que *Profesores-Profesores Eméritos* (30.8%) es el grupo que más proporción de valores en el Q4 tiene, esto es, los perfiles que crecen por debajo del 1.02%. La segunda variable en orden de importancia es el área disciplinar. Esto permite precisar que los investigadores con los peores ratios de crecimiento son *Profesores-Profesores Eméritos* de *Ciencias de la Vida-Arte y Humanidades* (38.3%) y *Ciencias Sociales-Ciencias de la Salud-Multidisciplinar* (34.6%). Por otro lado, *Profesores Asistentes* de *Ciencias Sociales-Arte y Humanidades* (44.7%) son los científicos que más incrementan su impacto.

Finalmente, la Figura 4 representa el árbol de decisión para la evolución del índice h. De forma similar a la Figura 3, *Ayudantes de Investigación* (80.8%) y *Becarios pre-doctorales* (68.6%) son los autores que tienen una mayor proporción de casos en Q1 y Q2, lo que significa que estos son los puestos académicos que más incrementan sus índices h. Contrariamente, *Profesores-Profesores Eméritos* (67.6%) y *Profesores Asociados* (45.1%) son el segmento que mayor proporción de casos en Q3 y Q4, siendo las escalas académicas que menos incrementan sus índices h. Sólo en el caso de *Profesor Asistente* se encontraron diferencias disciplinarias, aunque estas no fueran muy sustanciales. De este modo, *Ciencias Sociales-Arte y Humanidades-Multidisciplinar* contienen la mayor proporción de autores en Q1 (39.8%), mientras *Profesor Asistente* de *Ciencias de la Vida-Ciencias Físicas-Ciencias de la Salud* ocupan sus casos en Q2 y Q3 (58%). *Profesores-Profesores Eméritos* y *Profesores Asociados* son las únicas escalas académicas donde se apreciaron diferencias de género. Así, mujeres *Profesores-Profesores Eméritos* levemente muestran un mayor incremento de índice h que en los hombres con un 38.5% en Q1 y Q2, diferente al 31.5% de los hombres. Estas diferencias son más marcadas en el caso de *Profesores Asociados* donde las mujeres obtienen un 67.4% en Q1 y Q2, mientras que los hombres sólo alcanzan un 52.3%.

6. DISCUSIÓN

El uso de perfiles de Google Scholar Citations hace posible analizar directamente la actividad de los autores sin agrupar sus artículos, evitando los bien conocidos problemas de desambiguación y asignación de publicaciones (Wooding y otros, 2006; D'Angelo y otros, 2011). En el caso de este servicio, es el mismo autor el que crea el perfil, añadiendo, eliminando y uniendo sus publicaciones. Esto asegura una alta fiabilidad de los perfiles ya que estas publicaciones corresponden verdaderamente a dichos autores y no a otros con nombres similares. Otra ventaja es que estas publicaciones vienen de la base de datos de Google Scholar la cual está considerada el buscador científico más

Tabla II. Distribución del porcentaje de crecimiento de Cit./Art. e índice h, agrupado por cuartiles

Cuartiles	*Cit./Art.*	*N*	*Índice h*	*N*
Q1	5.87 - 133.99	759	4.8 - 51.67	760
Q2	2.99 - 5.86	759	2.83 - 4.75	775
Q3	1.03 - 2.98	758	1.61 - 2.78	741
Q4	0 - 1.02	758	0 - 1.6	758
Total		3034		3034

Figura 3. Árbol de decisión en función del porcentaje de crecimiento de Cit./Art., agrupado en cuartiles

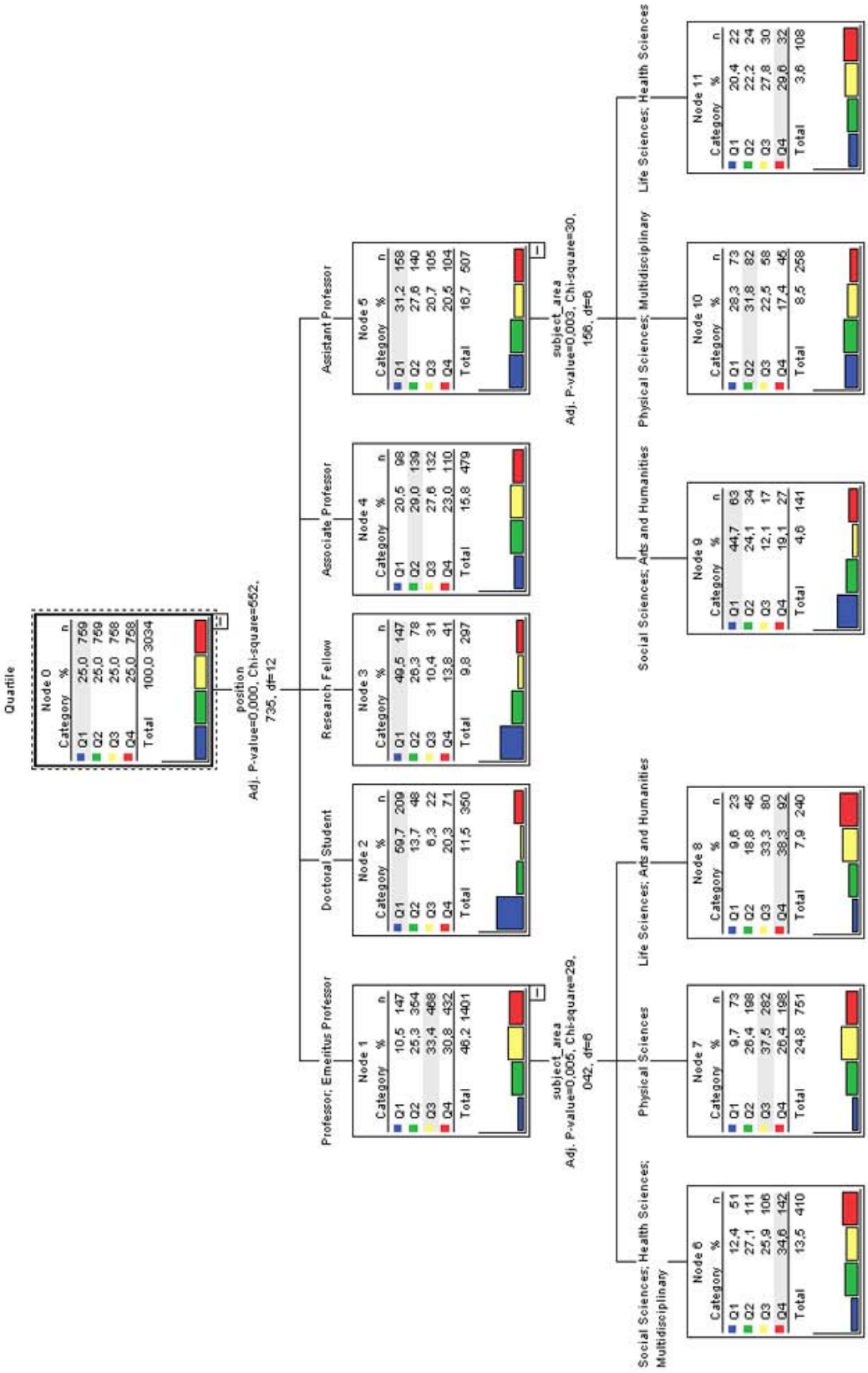
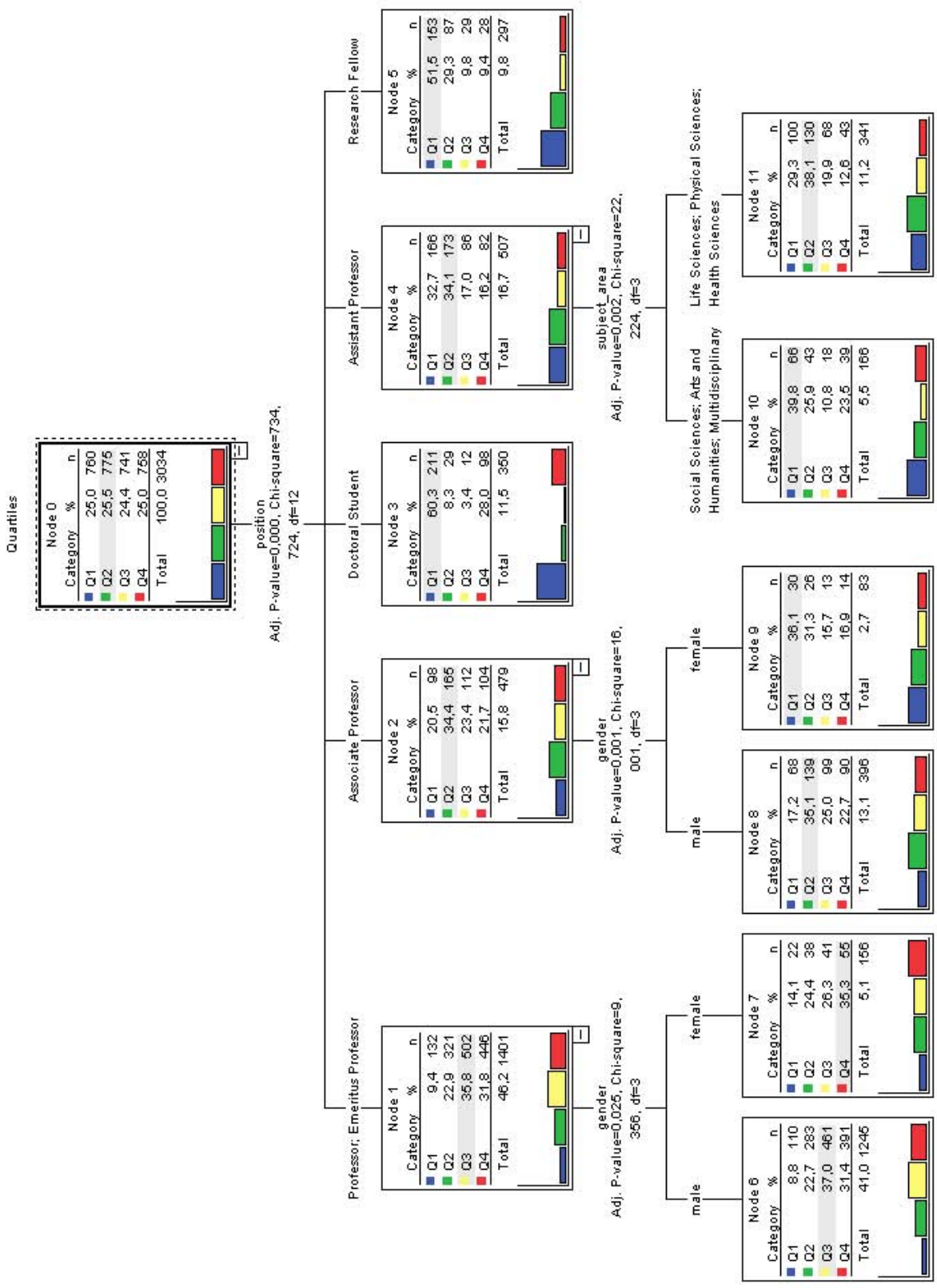


Figura 4. Árbol de decisión de acuerdo al porcentaje de incremento de índice h clasificado en cuartiles



 Rev. Esp. Doc. Cient., 38(4), octubre-diciembre 2015, e102. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2015.4.1225

exhaustivo (Meho y Yang, 2007; Kousha y Thelwall, 2007). En consecuencia estos resultados son un reflejo amplio y completo del impacto científico de estos investigadores. Sin embargo, este servicio de perfiles presenta algunas limitaciones que deben ser tenidas en consideración. Por ejemplo, es posible que la información que muestren los perfiles no sea correcta o esté desactualizada. Este problema es mínimo, ya que precisamente el hecho de que el autor cree el perfil hace que esta información sea la más veraz posible. Otros estudios han informado sobre una proporción significativa de citas asignadas a artículos equivocados (Bar-Ilan, 2008; García-Pérez, 2010) lo cual podría alterar los indicadores bibliométricos de un perfil. En este estudio, estos errores fueron considerados un problema horizontal y por lo tanto estos influyen por igual a cada grupo estudiado (materias, puestos y género). Otro problema que sí podría introducir un sesgo en las muestras es que Google Scholar presenta una pobre cobertura de materiales previos a 1980 (Pauly y Stergiou, 2005). Esto podría afectar a autores veteranos, principalmente *Profesores* y *Profesores Eméritos* con una trayectoria anterior a 1980, lo cuales podrían verse infrarrepresentados. Aunque quizás el problema más serio pueda ser la posibilidad de manipular un perfil a través de la subida de falsos artículos llenos de autocitaciones a repositorios no supervisados (Delgado López-Cózar y otros, 2014). Sin embargo, se puede considerar que el número de perfiles manipulados podrían ser muy pequeño al ser estas actitudes de dudosa ética científica. Por ejemplo, en esta muestra, el mayor incremento de índice h de un autor es 51.6%, y sólo nueve investigadores incrementan su ratio de Cit./Art. por encima del 50% en dos años. Estos números no evidencian conductas no éticas y podemos afirmar que la manipulación de perfiles es un hecho excepcional. En general, el uso de perfiles de Google Scholar Citations puede ser considerada una herramienta recomendable para análisis bibliométricos al nivel de autor ya que permite de forma fácil y pormenorizada seguir la producción e impacto científico de un gran número de investigadores.

Antes de los resultados en sí, es interesante también apreciar la ausencia de mujeres en la muestra. Sólo el 14.8% de los autores son mujeres, un porcentaje inferior a otras estadísticas (20-25%) (NSF, 2013; Landivar, 2013). La razón de esta diferencia puede ser debido a la alta presencia de perfiles de países emergentes (Brazil, India) con un menor presencia de la mujer en actividades científicas (Larivière y otros, 2013). Esta menor presencia de mujeres podría infravalorar la importancia del género a la hora de diferenciar la actividad científica, pero anteriores resultados confirman que no existen diferencias significativas entre hombres y mujeres cuando son citados (Ding y otros, 2006; Penas y Willett, 2006).

Con respecto a las diferencias entre puestos académicos y grupos temáticos, los resultados muestran que el primer elemento para distinguir los resultados científicos es la escala académica. En este sentido, se podría afirmar que la carrera es el factor que más influye en el éxito bibliométrico de un autor (Penner y otros, 2013). Así, jóvenes investigadores con una carrera inicial describen resultados por debajo a investigadores consolidados con una larga carrera como son los *Profesores y Profesores Eméritos*. Estos resultados están en consonancia con análisis anteriores (Ventura y Mombrú, 2006; Abramo y otros, 2009; Pagel y Hudetz, 2011) y son explicados generalmente como un fenómenos de ventaja acumulativa (Cole and Cole, 1972; Long, 1978). Sin embargo, esta situación cambia cuando el crecimiento es considerado. En este caso, los *Becarios pre-doctorales y Ayudantes de Investigación* son aquellos que más incrementan sus curriculums, mientras que los *Profesores* mantienen estable sus perfiles con ligeros crecimientos. Este hecho puede ser un reflejo de un fenómeno evolutivo en el que pequeñas entidades crecen más rápido que las grandes (Gibrat, 1930), haciendo que muchos investigadores noveles desarrollen sus curriculums en sus fases tempranas, experimentando crecimiento iniciales de gran importancia que marcarán el futuro de su prestigio (Maranto y Streuly, 1994).

Los resultados señalan que el segundo factor en importancia para diferenciar el impacto científico de un investigador es la disciplina científica. En general, se aprecia que los investigadores de *Arte y Humanidades* y *Ciencias Sociales* tienen un menor impacto que investigadores de *Ciencias de la Vida*, la disciplina que más autores sitúa en Q1. Este resultado se ajusta a anteriores análisis donde las biociencias consiguen más citas por artículo que otras disciplinas (Radicchi y otros, 2008), y donde las humanidades son escasamente citadas (Althouse y otros, 2009). Sin embargo, en función de la evolución de los indicadores por áreas temáticas, se aprecia un patrón interesante. Cuando el modelo detecta grupos por afinidad temática en la categoría de *Profesores-Profesores Eméritos*, estos grupos describen una conducta similar, concluyendo que no existen diferencias significativas entre disciplinas dentro de esta categoría profesional. Lo que sugiere que los investigadores sénior ralentizan sus carreras independientemente de los campos de investigación. Sin embargo, los investigadores más jóvenes sí describen diferentes patrones de crecimiento en función del área de investigación. Así,

investigadores noveles, especialmente *Profesores Asistentes*, de *Arte y Humanidades* y *Ciencias Sociales* experimentan mucho mayor incremento que sus colegas de *Ciencias de la Vida* y *Ciencias de la Salud*. Como en el caso de los puestos académicos, es posible que estas diferencias temáticas sean debidas a fenómenos de crecimiento donde los investigadores con una pequeña actividad incrementen más rápido su impacto científico que los autores con una gran producción. En cierto modo, se podría afirmar que los *Profesores Asistentes* de *Ciencias de la Vida* y *Ciencias de la Salud* consiguen la madurez científica más temprano que sus colegas de *Arte y Humanidades* y *Ciencias Sociales*; y la rápida ralentización en la actividad de los investigadores de *Ciencias de la Vida* y *Ciencias de la Salud* puede ser entendido como un signo de estabilidad, mientras que los *Profesores Asistentes* de *Arte y Humanidades* y *Ciencias Sociales* están aún desarrollando sus carreras (Smeby, 1998). En este sentido se podría sugerir que el esfuerzo en consolidar una carrera in Ciencias Sociales y Humanidades podría ser mayor que en el resto de áreas de investigación.

7. CONCLUSIONES

Los árboles de decisión han permitido concluir que el primer aspecto cualitativo que diferencia la actividad científica en términos de impacto es el puesto académico de un autor. Investigadores con una carrera establecida obtienen de este modo mejor impacto científico que los investigadores iniciados como consecuencia de una ventaja acumulativa. Esta influencia es también observada en función de disciplinas, detectando que los autores de ciencias de la vida consiguen más impacto que los investigadores de artes y humanidades. Sin embargo, los resultados no encuentran diferencias significativas entre hombres y mujeres en conseguir impacto.

Desde un punto de vista evolutivo, los árboles de decisión muestran que los investigadores jóvenes, principalmente de humanidades y ciencias sociales, incrementan sus curriculums más rápido que lo profesores sénior los cuales describen pequeños incrementos en todas las áreas estudiadas. Esto podría ser interpretado como un fenómeno de crecimiento en el que los investigadores noveles tienden a incrementar sus curricula en las fases iniciales de sus carreras para luego permanecer estables en su madurez.

Los árboles de decisión también han permitido agrupar y categorizar qué tipo de autores describen un mayor impacto científico considerando su género, puesto y disciplina tanto de una forma estática como longitudinal. Por lo tanto se puede concluir que esta herramienta de la minería de datos es recomendable para estudiar la influencia de varios aspectos cualitativos implicados en la actividad científica en relación con en el impacto y la producción, mostrando qué elementos de un perfil condicionan en mayor o menor manera una carrera prometedora.

Por último, Google Scholar Citations puede ser valorado como una apropiada herramienta bibliométrica porque facilita la construcción de exhaustivos y actualizados perfiles de autores con indicadores bibliométricos comparables entre sí. Sin embargo, se recomienda un proceso de limpieza previo de estos datos para evitar perfiles duplicados y manipulados, además de normalizar afiliaciones y nombres.

8. REFERENCIAS

Abramo, G.; D'Angelo, C. A.; Caprasecca, A. (2009). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics*, vol. 79(3), 517-539. http://dx.doi.org/10.1007/s11192-007-2046-8

Aguillo, I. F.; Ganadino, B.; Ortega, J. L.; Prieto, J. A. (2005). What the Internet says about Science. *Scientist*, vol. 19(14), 10-11.

Aksnes, D. W.; Rorstad, K.; Piro, F.; Sivertsen, G. (2011). Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology,* vol. 62(4), 628–636. http://dx.doi.org/10.1002/asi.21486

Althouse, B. M.; West, J. D.; Bergstrom, C. T.; Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology,* vol. 60(1), 27–34. http://dx.doi.org/10.1002/asi.20936

Amaral, L. A. N.; Scala, A.; Barthelemy, M.; Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, vol. 97(21), 11149-11152. http://dx.doi.org/10.1073/pnas.200327197

Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, vol. 74(2), 257-271. http://dx.doi.org/10.1007/s11192-008-0216-y

Chen, X. (2010). Google Scholar's Dramatic Coverage Improvement Five Years after Debut. *Serials Review*, vol. 36(4), 221-226. http://dx.doi.org/10.1016/j.serrev.2010.08.002 / http://dx.doi.org/10.1080/00987913.2010.10765321

Cole, J. R.; Cole, S. (1972). The Ortega Hypothesis. *Science*, vol. 178(October 27), 368-375.

D'Angelo, C. A.; Giuffrida, C.; Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, vol. 62(2), 257-269. http://dx.doi.org/10.1002/asi.21460

Delgado López-Cózar, E.; Robinson-García, N.; Torres-Salinas, D. (2014), The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, vol. 65(3), 446–454. http://dx.doi.org/10.1002/asi.23056

Ding, W. W.; Murray, F.; Stuart, T. E. (2006). Gender differences in patenting in the academic life sciences. *Science*, vol. 313, 665–667. http://dx.doi.org/10.2139/ssrn.1260388 / http://dx.doi.org/10.1126/science.1124832

Finardi, U. (2014). On the time evolution of received citations, in different scientific fields: An empirical study. *Journal of Informetrics*, vol. 8(1), 13-24. http://dx.doi.org/10.1016/j.joi.2013.10.003

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the American Society for Information Science and Technology*, vol. 61(10), 2070-2085. http://dx.doi.org/10.1002/asi.21372

Garfield, E. (1980). The Number of Biochemical Articles Is Growing, But Why Also the Number of References per Article? *Essays of an Information Scientist,* vol. 4, 414-418.

Gibbons, M.; Limoges, C.; Nowotny, H.; Schwartzman, S.; Scott, P.; Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London; Sage.

Gibrat, R. (1931). *Les Inégalités économiques.* Paris; Recueil Sirey.

Hancock, T.; Lane, J.; Ray, R.; Glennon, D. (1992). The ombudsman: factors influencing academic research productivity: a survey of management scientists. *Interfaces*, vol. 22(5), 26-38. http://dx.doi.org/10.1287/inte.22.5.26

Huang, Z.; Yuan, B. (2012). Mining Google Scholar Citations: An Exploratory Study. *Lecture Notes in Computer Science*, vol. 7389, 182-189. http://dx.doi.org/10.1007/978-3-642-31588-6_24

Jacobs, D.; Ingwersen, P. (2000). A bibliometric study of the publication patterns in the sciences of South African scholars 1981–96. *Scientometrics*, vol. 47(1), 75-93. http://dx.doi.org/10.1023/A:1005617825947

Kidd, J. S. (1988). The popularization of science: Some basic measurements. *Scientometrics*, vol. 14(1), 127-142. http://dx.doi.org/10.1007/BF02020247

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, vol. 43(1), 27-43. http://dx.doi.org/10.1007/BF02458392

Kousha, K.; Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multidiscipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, vol. 58(7), 1055-1065. http://dx.doi.org/10.1002/asi.20584

Kyvik, S.; Teigen, M. (1996). Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology & Human Values*, vol. 21(1), 54-71. http://dx.doi.org/10.1177/016224399602100103

Landivar, L. C. (2013). *Disparities in STEM Employment by Sex, Race, and Hispanic Origin*. American Community Survey Reports, ACS-24, [17 de septiembre de 2014] http://www.census.gov/prod/2013pubs/acs-24.pdf

Larivière, V.; Ni, C.; Gingras, Y.; Cronin, B.; Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, vol. 504(12 December 2013), 211–213. http://dx.doi.org/10.1038/504211a

Long, J. S. (1978). Productivity and academic position in the scientific career. *American Sociological Review*, vol. 43(6), 889-908. http://dx.doi.org/10.2307/2094628

Long, J. S. (2001). *From scarcity to visibility: Gender differences in the careers of doctoral scientists and engineers*. Washington, DC; National Academies Press.

Maranto, C. L.; Streuly, C. A. (1994). The Determinants of Accounting Professors' Publishing Productivity—The Early Career. *Contemporary Accounting Research,* vol. 10(2), 387–407. http://dx.doi.org/10.1111/j.1911-3846.1994.tb00399.x

Meho, L. I.; Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty:

Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, vol. 58(13), 2105-2125. http://dx.doi.org/10.1002/asi.20677

National Science Foundation. (2013). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013*. Special Report NSF 13-304. Arlington, VA. [17 de septiembre de 2014] http://www.nsf.gov/statistics/wmpd

Orduña-Malea, E.; Delgado López-Cózar, E. (2014). Google Scholar Metrics evolution: an analysis according to languages. *Scientometrics*, vol. 98(3), 2353-2367. http://dx.doi.org/10.1007/s11192-013-1164-8

Ortega, J. L. (2015). How is a scientific information web service settled? A demographic study of Google Scholar Citations population. *Scientometrics*, vol. 104(1), 1-18. http://dx.doi.org/10.1007/s11192-015-1593-7

Ortega, J. L.; Aguillo, I. F. (2012). Science is all in the eye of the beholder: keyword maps in Google Scholar Citations. *Journal of the American Society for Information Science and Technology,* vol. 63(12), 2370-2377. http://dx.doi.org/10.1002/asi.22761

Ortega, J. L.; Aguillo, I. F. (2013). Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations. *Journal of Informetrics,* vol. 7(2), 394-403.

Pagel, P. S.; Hudetz, J. A. (2011). An analysis of scholarly productivity in United States academic anaesthesiologists by citation bibliometrics. *Anaesthesia,* vol. 66(10), 873–878. http://dx.doi.org/10.1111/j.1365-2044.2011.06860.x

Pauly, D.; Stergiou, K. I. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, vol. 2005, 33-35.

Peñas, C. S.; Willett, P. (2006). Brief communication: Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, vol. 32(5), 480-485. http://dx.doi.org/10.1177/0165551506066058

Penner, O.; Pan, R. K.; Petersen, A. M.; Kaski, K.; Fortunato, S. (2013). On the predictability of future impact in science. *Scientific reports*, vol. 3. http://dx.doi.org/10.1038/srep03052

Pitney, W. A.; Gilson, T. A. (2012). Educational technology: Using Google Scholar Citations to support the impact of scholarly work. *Athletic Training Education Journal,* vol. 7(1), 38-39. http://dx.doi.org/10.5608/070138

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, vol. 493(7431), 159-159.

Radicchi, F.; Fortunato, S.; Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, vol. 105(45), 17268-17272. http://dx.doi.org/10.1073/pnas.0806977105

Schubert, A.; Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, vol. 9(5-6), 281-291. http://dx.doi.org/10.1007/BF02017249

Scopus (2014). Subject Area Categories. [17 de septiembre de 2014] http://help.scopus.com/Content/h_subject_categories.htm

Small, H. G.; Crane, D. (1979). Specialties and disciplines in science and social science: an examination of their structure using citation indexes. *Scientometrics*, vol. 1(5-6), 445-461. http://dx.doi.org/10.1007/BF02016661

Small, H.; Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies,* vol. 4, 17-40. http://dx.doi.org/10.1177/030631277400400102

Smart, J. C.; Bayer, A. E. (1986). Author collaboration and impact: A note on citation rates of single and multiple authored articles. *Scientometrics*, vol. 10(5), 297-305. http://dx.doi.org/10.1007/BF02016776

Smeby, J. C. (1998). Knowledge production and knowledge transmission. The interaction between research and teaching at universities. *Teaching in Higher Education*, vol. 3(1), 5-20. http://dx.doi.org/10.1080/1356215980030101

Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and nonscience. En: Carnot, E.N.; Pollack, D. *(*editores*).Communication among scientists and engineers*, Lexington, D.C; Heath Lexington Books and Company, pp. 3-22.

Ventura, O. N.; Mombrú, A. W. (2006). Use of bibliometric information to assist research policy making. A comparison of publication and citation profiles of Full and Associate Professors at a School of Chemistry in Uruguay. *Scientometrics*, vol. 69(2), 287-313. http://dx.doi.org/10.1007/s11192-006-0154-5

Wooding, S.; Wilcox-Jay, K.; Lewison, G.; Grant, J. (2006). Co-author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. *Scientometrics*, vol. 66(1), 11-21. http://dx.doi.org/10.1007/s11192-006-0002-7