

Bibliometric Impact Assessment With R and the CITAN Package

Marek Gagolewski^{a,b}

E-mail: gagolews@ibspan.waw.pl

^a*Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland*

^b*Faculty of Mathematics and Information Science, Warsaw University of Technology
pl. Politechniki 1, 00-661 Warsaw, Poland*

Abstract

In this paper **CITAN**, the CITation ANalysis package for R statistical computing environment, is introduced. The main aim of the software is to support bibliometricians with a tool for preprocessing and cleaning bibliographic data retrieved from *SciVerse Scopus* and for calculating the most popular indices of scientific impact.

To show the practical usability of the package, an exemplary assessment of authors publishing in the fields of scientometrics and webometrics is performed.

Keywords: data analysis software, quality control in science, citation analysis, bibliometrics, Hirsch's h index, Egghe's g index, *SciVerse Scopus*.

This is a revised version of the paper:

Gagolewski M., Bibliometric impact assessment with R and the CITAN package, *Journal of Informetrics*, 2011, doi:10.1016/j.joi.2011.06.006.

1. Introduction

The introduction of the h -index by J.E. Hirsch (2005) started a very intensive research trend in the field of scientometrics. Numerous bibliometric impact indices, like the g -index (Egghe, 2006b), the w -index (Woeginger, 2008b), or the R -index (Jin et al., 2007) are particular instances of a wide class of functions called aggregation operators (cf. Gagolewski and Grzegorzewski, 2010, 2011a,b). Such operators merge several numerical values into a single, representative one. They may be applied in many areas like engineering, statistics, economy or social sciences. The theory of aggregation is a rapidly developing mathematical domain, see (Grabisch et al., 2009) for the recent state of the art monograph.

Operators that aggregate the number of citations received by scientific publications may support fair assessment of researchers and therefore be conceived as tools for the quality control in science. Moreover, bibliometrics uses them as tools describing the current state of various science-related domains of human activity (cf. e.g. Franceschini and Maisano, 2011; Norris and Oppenheim, 2010).

In practice, we gather citation data from on-line bibliographic databases. Such services often base on information received directly from the publishers. However, as the amount of data is huge, they must use automatic or at least semi-automatic processing. This may lead to a number of problems like: incorrectly mapped citations, non-unique author or document representation, etc. Therefore one need to be conscious that such data are uncertain and imprecise in its very nature (cf. Gagolewski and Grzegorzewski, 2011b). It is a great challenge for current computer science and applied mathematics to try to deal with these problems (d'Angelo et al., 2011; Baneyx, 2008).

According to *JISC Academic Database Assessment Tool*¹ and e.g. (Meho and Rogers, 2008; Meho and Sugimoto, 2009; Vieira and Gomes, 2009; Moed, 2010; González-Pereira et al., 2010) Elsevier's *SciVerse Scopus* is a general bibliographic database that covers the largest number of peer-reviewed sources. However, even though it gives citation data for the documents, and is able to export a few thousand records at once to a number of file formats, it is far less often used in bibliometric analyses than the currently most popular Thomson's *Web of Science*. Among important reasons for this situation we find lack of software for processing and cleaning publication information

¹<http://www.jisc-adat.com/adat>

adapted for such output.

Moreover, regardless of the usage of statistical methods in this type of research, we do not find any powerful libraries *integrated into* statistical data analysis software. Such a tool could support scientometricians in performing their analyses.

The main aim of this paper is therefore to introduce the **CITAN** package written in a highly programmable, platform-independent language for statistical computing, **R**, which tries to overcome the two above-mentioned problems. The presentation of software features is given in such a way that it forms a step-by-step usage guide for the practitioners. We show how it can be used to assess the authors by means of the most interesting indices of scientific impact.

Unlike existing software, for example *Publish or Perish* (Harzing and van der Wal, 2008) together with CleanPoP (Baneyx, 2008), the open architecture of **CITAN** easily allows for adding and testing new functionality. What is more, it is designed not only to process one author's record, but also many records at once.

The paper is organized as follows. In Sec. 2 we present some elementary facts about *Scopus*. Section 3 describes the most basic features of the **CITAN** package. Also, a brief review of popular bibliometric impact indices is given. In Sec. 4 we show how the software may be used to perform the assessment of researchers publishing in scientometrics and webometrics. Section 5 summarizes the paper.

2. *SciVerse Scopus*

Let us first present some basic facts on *SciVerse Scopus* as this database will be used as a source of bibliographic information in our software package².

The first version of the Elsevier's product has been launched in November, 2004. According to JISC ADAT, on April 24, 2011 *Scopus* was indexing 30,017 sources (i.e. scientific journals and conference proceedings with ISSNs), of which 18,772 were active titles, from ca. 5000 international publishers (e.g. Elsevier, Springer-Verlag, Kluwer Academic Publishers, John

²If not stated explicitly, information is according to <http://www.info.sciverse.com/scopus/scopus-in-detail/facts/>, last updated June 2010.

Wiley & Sons, IEEE). This gave a total of 44,477,066 documents (including 3.3 mil. conference papers). Generally, about 2 mil. new records are added each year. *Scopus* has currently the largest source coverage of the general, peer-reviewed bibliographic databases available (Meho and Rogers, 2008; Meho and Sugimoto, 2009; Vieira and Gomes, 2009).

Scopus indexes sources from all geographic regions. It also includes publications whose main language is not English (21% overall). The majority of the documents published after 1996 have full bibliographies. However, as citation information is automatically generated, we will see below that it is likely to contain errors. The most cited article has 90,933 citations³.

2.1. ASJC: Classification of sources

Each source is given one or more four-digit ASJC (*All Science: Journal Classification*) classification codes, which may be used to determine its main subject-matter.

A full list of top-level codes is presented in Tab. 1. Many of them are divided into subcategories. For example, the top-level category 1700 (Computer Science) includes the following: 1701 (Miscellaneous), 1702 (Artificial Intelligence), 1703 (Computational Theory and Mathematics), 1704 (Computer Graphics and Computer-Aided Design), 1705 (Computer Networks and Communications), 1706 (Computer Science Applications), 1707 (Computer Vision and Pattern Recognition), 1708 (Hardware and Architecture), 1709 (Human-Computer Interaction), 1710 (Information Systems), 1711 (Signal Processing), and 1712 (Software).

As we will see further on, the codes may also be used to classify individual publications.

³Laemmli, U.K. (1970), Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature* **227**(525), 680–685.

Table 1: Top-level ASJC codes.

ASJC	Description
1000	General
1100	Agricultural and Biological Sciences
1200	Arts and Humanities
1300	Biochemistry, Genetics and Molecular Biology
1400	Business, Management and Accounting
1500	Chemical Engineering
1600	Chemistry
1700	Computer Science
1800	Decision Sciences
1900	Earth and Planetary Sciences
2000	Economics, Econometrics and Finance
2100	Energy
2200	Engineering
2300	Environmental Science
2400	Immunology and Microbiology
2500	Materials Science
2600	Mathematics
2700	Medicine
2800	Neuroscience
2900	Nursing
3000	Pharmacology, Toxicology and Pharmaceutics
3100	Physics and Astronomy
3300	Social Sciences
3400	Veterinary
3500	Dentistry
3600	Health Professions

2.2. Types of documents

In bibliometric analyses it is often very important to take the types of documents into account. For example, sometimes we want to base our surveys only on peer-reviewed publications as they — at least theoretically — should be of higher quality.

Here is a list of document types used in *Scopus*:

- **Article (ar)** — presents original research or opinion, also a post-conference paper; in peer-reviewed journals; usually several pages in length.
- **Article-in-Press (ip)** — an article available on-line before official publication.
- **Conference paper (cp)** — a publication presented at a conference or symposium.
- **Editorial (ed)** — summarizes several articles or provides editorial news.
- **Erratum (er)** — reports an error of a previously published paper in the same journal.
- **Letter (le)** — correspondence with the editor.
- **Note (no)** — a short note, discussion or commentary.
- **Review (re)** — significant review of original research; often has extensive bibliography.
- **Short survey (sh)** — a short review of original research; shorter in length than a **Review**.

In next section we introduce the **CITAN** package. Also, we review the most interesting aggregation operators that may be used to assess the merit of individuals.

3. The CITAN package

3.1. Package description and installation

R (R Development Core Team, 2011) is a very popular, freely available language and environment for statistical computing and graphics. It implements a large variety of statistical tools, including linear and nonlinear models, time series, classification, clustering, microarray analysis, etc. For more information on the software the reader is encouraged to visit its homepage at <http://www.r-project.org/about.html>, or refer e.g. to (Crawley, 2007).

The **CITAN** package (Gagolewski, 2011) is an open source (GNU LGPL v3 license) toolpack for citation analysis. It requires R version 2.10 or higher

and a few other packages like **RSQLite** (James, 2010) for an interface to SQLite (a flat file-based relational database management system), or **RGtk2** (Lawrence and Temple Lang, 2010) which allows for creating the graphical user interface.

CITAN is available via the CRAN (The Comprehensive R Archive Network) repository, therefore it may be downloaded and installed directly from R, just as any other officially maintained package. During the installation process all dependencies should be automatically resolved. The package has been tested under all major operating systems including Windows, Linux and MacOS.

R version 2.13.0 (2011-04-13)

Copyright (C) 2011 The R Foundation for Statistical Computing

ISBN 3-900051-07-0

Platform: x86_64-redhat-linux-gnu (64-bit)

```
> install.packages("CITAN"); # download and install the package
```

```
> library("CITAN"); # load the package
```

```
Loading required package: RSQLite
```

```
Loading required package: RGtk2
```

```
...
```

Each element of the package has been documented. We may access the user's manual by calling the following command.

```
> ?CITAN # view the manual
```

```
CITAN-package
```

```
package:CITAN
```

```
R Documentation
```

```
CITation ANalysis toolpack
```

```
...
```

```
Author(s):
```

```
Marek Gagolewski <gagolews@ibspan.waw.pl>
```

In this paper we present the most important package features so that the reader will be able to conduct his/her own bibliometric surveys. **CITAN** in version 2011.05-1 is used for the illustration. However, please note that the package is constantly being improved and many new functions are planned to be added in the future.

Let us first take a look at available impact functions.

3.2. Impact functions

Let \mathbb{N}_0 denote the set of nonnegative integers and $\mathbb{N}_0^{1,2,\dots} = \bigcup_{n=1}^{\infty} \mathbb{N}_0^n$ the set of all arbitrary-length sequences of elements in \mathbb{N}_0 . In our context, each $\mathbf{x} \in \mathbb{N}_0^{1,2,\dots}$ represents an author's citation record.

Definition 1. An **impact function** (Gagolewski and Grzegorzewski, 2010, 2011a,b) in $\mathbb{N}_0^{1,2,\dots}$ is an aggregation operator $J : \mathbb{N}_0^{1,2,\dots} \rightarrow \mathbb{R}$ which:

1. is nondecreasing in each variable:

$$(\forall n)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{N}_0^n) \quad \mathbf{x} \leq \mathbf{y} \Rightarrow J(\mathbf{x}) \leq J(\mathbf{y}),$$

2. is arity-monotonic, i.e. $(\forall n, m)(\forall \mathbf{x} \in \mathbb{N}_0^n)(\forall \mathbf{z} \in \mathbb{N}_0^m) \quad J(\mathbf{x}) \leq J(\mathbf{x}, \mathbf{z})$,
3. is symmetric, i.e. $(\forall n)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{N}_0^n) \quad \mathbf{x} \cong \mathbf{y} \Rightarrow J(\mathbf{x}) = J(\mathbf{y})$,
4. fulfills the weak lower boundary condition: $\inf_{\mathbf{x} \in \mathbb{N}_0^{1,2,\dots}} J(\mathbf{x}) = 0$,
5. fulfills the weak upper boundary condition: $\sup_{\mathbf{x} \in \mathbb{N}_0^{1,2,\dots}} J(\mathbf{x}) = \infty$,

where for $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{z} = (z_1, \dots, z_m)$, we write $\mathbf{x} \leq \mathbf{y}$ iff $x_1 \leq y_1, x_2 \leq y_2, \dots$ and $x_n \leq y_n$, $(\mathbf{x}, \mathbf{z}) = (x_1, \dots, x_n, z_1, \dots, z_m) \in \mathbb{N}_0^{n+m}$, and $\mathbf{x} \cong \mathbf{y}$ iff there exists a permutation σ such that $\mathbf{x} = (y_{\sigma(1)}, \dots, y_{\sigma(n)})$.

Such a subclass of aggregation operators (see Grabisch et al., 2009) may be used to assess an author's scientific output by means of citation counts received by his/her individual publications. A similar axiomatization for bibliometric impact indices was proposed by Woeginger (2008a,b). Moreover, it is worth noting that the proposed class of functions may be used in many other domains. See (Gagolewski and Grzegorzewski, 2010, 2011a,b) and also (Franceschini and Maisano, 2009) for more exemplary applications.

We will now recall some popular impact functions appearing in the literature. Each of them may be easily calculated with **R** and the **CITAN** package.

3.2.1. The *h-index* and its variants

Traditionally, scientometricians had used a few simple impact functions to assess a researcher's output. Among them we have the author's productivity ($\text{length}(\mathbf{x}) = |\mathbf{x}|$), the overall citation count ($\text{sum}(\mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} x_i$) or the maximal citation count ($\text{max}(\mathbf{x}) = \max_i x_i$). However, the start of intensive studies on bibliometric indices combining both the productivity of an author and the overall quality of the papers (see Glänzel, 2006) is due to J.E. Hirsch (2005), who proposed his so-called *h-index*. The introduction of the index had a great impact on the field; one of the researchers even divides the history of the subject into a pre-Hirsch and a post-Hirsch period (Prathap, 2010).

Definition 2. The *h-index* is an impact function `index.h` such that for $\mathbf{x} \in \mathbb{N}_0^n$ we have

$$\text{index.h}(\mathbf{x}) = \max \{i = 0, 1, \dots, n : x_{(n-i+1)} \geq i\}, \quad (1)$$

where $x_{(i)}$ denotes the i th order statistic, that is the i th smallest value of \mathbf{x} . Moreover, for the simplicity of notation we assume $x_{(n+1)} = x_{(n)}$.

Interestingly, a similar object has appeared earlier in the problem of Bonferroni-type multiple statistical significance testing, see e.g. (Benjamini and Hochberg, 1995).

Many simple modifications of the index have been proposed. However, some of them may be expressed as the *h-index* of a properly *transformed* input sequence. For example, the *h^α -index* (Kosmulski, 2006) for $\alpha \geq 1$ is defined as

$$\max \{i = 0, 1, \dots, n : x_{(n-i+1)} \geq i^\alpha\} = \text{index.h}(\mathbf{x}^{1/\alpha})$$

or the *h_β -index* (van Eck and Waltman, 2008; Wu, 2010), $\beta > 0$, is equal to

$$\max \{i = 0, 1, \dots, n : x_{(n-i+1)} \geq \beta i\} = \text{index.h}(\mathbf{x}/\beta),$$

where the vector exponentiation and division are understood element-wise here. Consequently, there is often no need to construct separate procedures to calculate such functions.

The computational method for determining the *h-index* works in $O(n)$ time on unsorted data (that is the number of required arithmetic operations is proportional to a linear function of the input sequence length) and is due to Woeginger (2010). The procedure bases on the “median-of-medians” selection algorithm (Blum et al., 1973, see also Beliaikov, 2011). Moreover, the $O(\log n)$ -time algorithm for pre-sorted data has been proposed in (Gagolewski and Grzegorzewski, 2009).

We may expect that the computational speedup gained by the usage of the latter procedure is insignificant in the case of small-sized samples. However, our package may also be used for meso- and macro-level bibliometric studies (e.g. assessing scientific institutes or whole countries), where such a performance improvement may be noticeable.

3.2.2. The g -index

Another popular impact function is the g -index proposed by L. Egghe (2006a,b).

Definition 3. The g -index is an impact function index.g such that for $\mathbf{x} \in \mathbb{N}_0^n$ we have

$$\text{index.g}(\mathbf{x}) = \max \left\{ i = 0, 1, \dots, n : \sum_{k=1}^i x_{(n-k+1)} \geq i^2 \right\}. \quad (2)$$

It is worth noting that this aggregation operator may also be expressed as the h -index of a function of the input vector. Nevertheless, the transformation is computationally far more complex here. We have $\text{index.g}(\mathbf{x}) = \text{index.h}(\sqrt{\Sigma \mathbf{x}})$, where $\Sigma \mathbf{x} = (x_{(n)}, x_{(n)} + x_{(n-1)}, \dots, x_{(n)} + x_{(n-1)} + \dots + x_{(1)})$.

For modifications of the g -index see e.g. (van Eck and Waltman, 2008; Guns and Rousseau, 2009; Alonso et al., 2010). An efficient, linear-time algorithm for computing this impact function was considered in (Woeginger, 2010).

3.2.3. The r_p -index

In (Gagolewski and Grzegorzewski, 2009) two interesting generalizations of the h -index were developed. The presentation of the **CITAN** package is a good occasion to empirically verify their potential practical usability.

Let $\mathbf{y} \in \mathbb{N}_0^n$ and $\mathbf{z} \in \mathbb{N}_0^m$. From now on, we write $\mathbf{y} \trianglelefteq \mathbf{z}$ iff $n \leq m$ and $y_{(n)} \leq z_{(m)}$, $y_{(n-1)} \leq z_{(m-1)}$, \dots , and $y_{(1)} \leq z_{(m-n+1)}$. We see that \trianglelefteq is a partial ordering of vectors in $\mathbb{N}_0^{1,2,\dots}$. Moreover, let $\lceil \cdot \rceil$ denote the ceiling function, i.e. the smallest integer not less than its argument.

The r_p -index may be expressed as follows.

Definition 4. The r_p -index for $p \geq 1$ is an impact function

$$\text{index.rp}(\mathbf{x}, p) := \sup \{ r > 0 : \mathbf{s}^{p,r} \trianglelefteq \mathbf{x} \}, \quad (3)$$

where $\mathbf{x} \in \mathbb{N}_0^n$ and $\mathbf{s}^{p,r} \in \mathbb{N}_0^{\lceil r \rceil}$, $r > 0$, denotes a sequence

$$\mathbf{s}^{p,r} = \begin{cases} (\sqrt[p]{r^p - 0^p}, \sqrt[p]{r^p - 1^p}, \dots, \sqrt[p]{r^p - \lceil r - 1 \rceil^p}) & \text{if } p < \infty, \\ (r, r, \dots, r) & \text{if } p = \infty. \end{cases} \quad (4)$$

It may be shown that the r_∞ -index is equivalent to the h -index and the r_1 -index is equal to the w -index proposed by Woeginger (2008b).

3.2.4. The l_p -index

The r_p -index may be extended as follows.

Definition 5. The l_p -index, $p \geq 1$, is a function that maps each $\mathbf{x} \in \mathbb{N}_0^{1,2,\dots}$ to a pair (a, b) , $a, b \geq 0$, such that

$$\text{index.lp}(\mathbf{x}, p) = \arg \sup_{(a,b)} \{ab : \mathbf{e}^{p,a,b} \trianglelefteq \mathbf{x}\}, \quad (5)$$

where $\mathbf{x} \in \mathbb{N}_0^n$ and $\mathbf{e}^{p,a,b} \in \mathbb{N}_0^{[a]}$, $a > 0$, $b > 0$, denotes a sequence

$$\mathbf{e}^{p,a,b} = \begin{cases} \left(\sqrt[p]{b^p - (\frac{b}{a}0)^p}, \sqrt[p]{b^p - (\frac{b}{a}1)^p}, \dots, \sqrt[p]{b^p - (\frac{b}{a}\lceil a-1 \rceil)^p} \right) & \text{if } p < \infty, \\ (b, b, \dots, b) & \text{if } p = \infty. \end{cases} \quad (6)$$

We see that the l_p -index is a function in $[0, \infty)^2$. Hence its usage in the impact assessment problem requires its projection to \mathbb{R} . It may easily be proved that e.g. $l_p^{(g)} := \sqrt{ab}$, where $\text{index.lp}(\mathbf{x}, p) = (a, b)$, is an impact function in $\mathbb{N}_0^{1,2,\dots}$.

Interestingly, if $\text{index.lp}(\mathbf{x}, \infty) = (a, b)$, then the value ab is equivalent to the MAXPROD-index proposed by Kosmulski (2007).

In (Gagolewski and Grzegorzewski, 2009) we have asked whether there exists an algorithm that may compute the value of the l_p -index for $p < \infty$ faster than in $O(n^3)$ time. Indeed, in an unpublished research report (Gagolewski et al., 2009), a procedure that runs in $O(n)$ time on ordered data was developed. It bases on the Graham's method for determining the convex hull of a planar set of points (Graham, 1972). Its implementation is of course included in our package.

3.3. Database model

For the purpose of storing bibliometric data for analysis we have designed the following relational database model (for alternative approaches see e.g. Yu et al., 2008; Mallig, 2010). It has been created so as to conform with the information generated by *Scopus*. However, its level of abstraction is quite high and thus support for other databases may easily be added in the future.

All analyzed data are stored in a so-called **local bibliometric storage** (LBS). It is currently maintained by the SQLite relational database management system (see <http://www.sqlite.org>). This light-weight RDBMS is automatically installed with the package and requires no configuration.

Fig. 1 presents an entity-relationship diagram for the database model. The reader interested in an SQL code used to create an LBS is referred to the manual page for the *lbsCreate()* function.

Among the most important objects represented in an LBS are: **sources** (e.g. journals, conference proceedings; identified by ISSNs), **documents** (e.g. articles, reviews; identified by *Scopus* eids extracted from URLs) and **authors** (identified by names).

Another significant feature of an LBS is the possibility of forming groups of documents. Such — non necessarily disjoint — collections are called **surveys**. As a consequence, there is no need to create separate LB storages for different bibliometric analyses as one document may belong to many groups.

Let us then create an LBS which will be used in the next section. First, we connect to an SQLite database, stored in the local file named **Scopus-biblio.db**.

```
> conn <- lbsConnect("Scopus-biblio.db"); # connect to the database
```

From now on, the **conn** object represents the established database connection and will be used to communicate with the RDBMS.

To create the above-described table scheme we call the following function.

```
> lbsCreate(conn); # create an LBS
```

```
Creating table 'Biblio_Categories'... DONE.  
Creating table 'Biblio_Countries'... DONE.  
Creating table 'Biblio_Sources'...  DONE.  
...
```

The process has been completed successfully. We are ready to perform an exemplary analysis of bibliometric data. We will also show how **CITAN** may be used to import and preprocess data retrieved from *Scopus*.

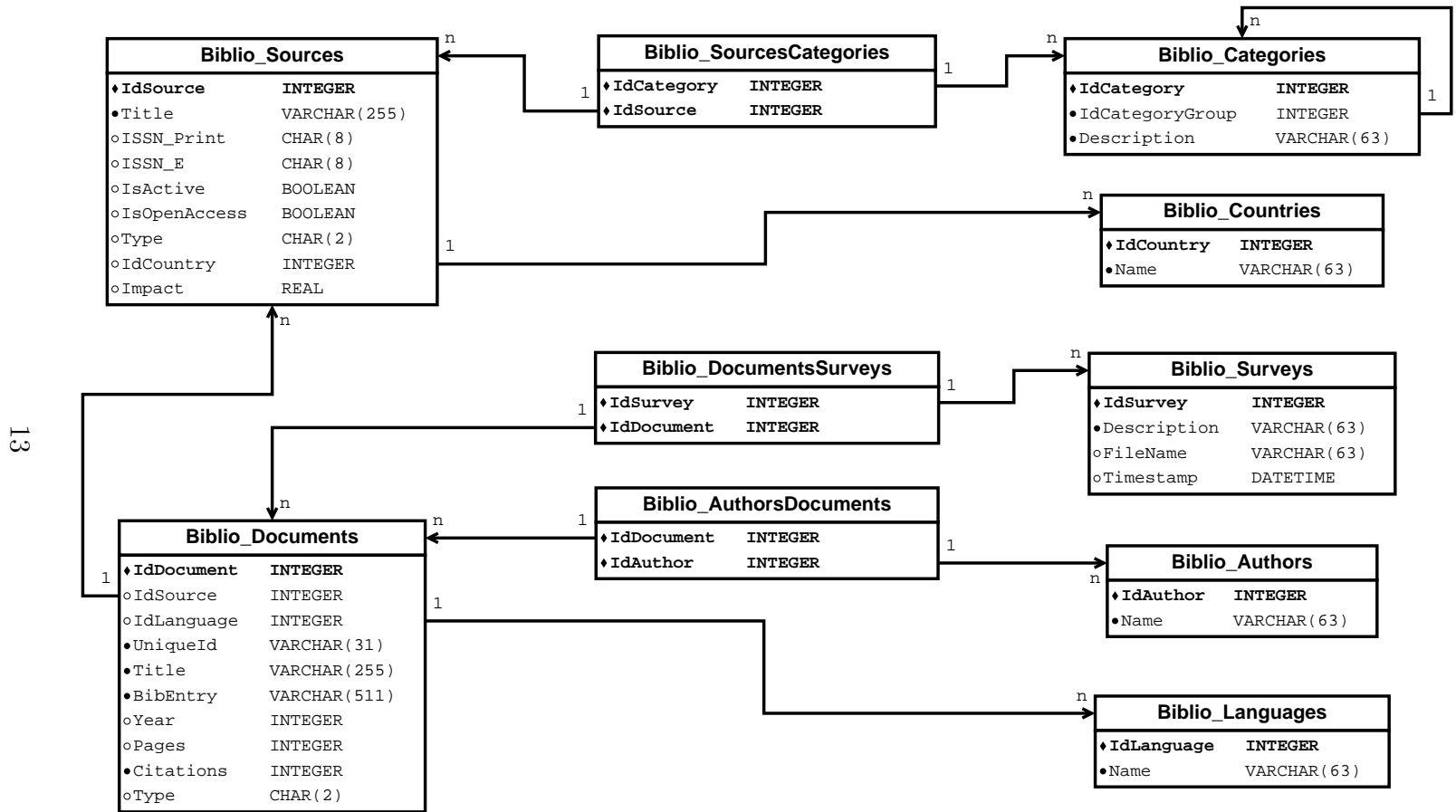


Figure 1: Local bibliometric storage model as created by the *lbsCreate()* function.

4. Exemplary analysis

Our sample consists of publications in scientometrics and webometrics. It is worth noting that we do not analyze the whole output of each scientist. Our main aim is therefore to indicate the authors of the greatest merit within the disciplines of concern. Such an investigation could be performed e.g. if we were to award a group of researchers a prize for their scientific achievements.

We used the following queries to extract publication data from *Scopus*:

1. TITLE(index quantify individual scientific research output) — 3 documents including the paper (Hirsch, 2005),
2. 795 documents citing the paper (Hirsch, 2005),
3. TITLE-ABS-KEY(bibliometric*) OR TITLE-ABS-KEY scientometric*) OR TITLE-ABS-KEY(webometric*) — 6870 documents; a keyword-based query,
4. SRCTITLE(Scientometrics) OR SRCTITLE(Informetrics) OR SRCTITLE(Cybermetrics) OR SRCTITLE("Annual Review of Information Science and Technology") OR SRCTITLE("Journal of the American Society for Information Science") — 5969 documents; a source-based query including the documents appearing in leading journals in the field like: *Scientometrics*, *Journal of Informetrics*, *Journal of the American Society for Information Science (and Technology)*.

The data were gathered on March 27, 2011, 13:33–14:05. We have accessed *Scopus* via the Main Library of Warsaw University of Technology (Biblioteka Główna Politechniki Warszawskiej).

Due to the fact that *Scopus* limits the number of exportable items to 2000 records, the queries had to be split into subgroups using the PUBYEAR directive. We used the following settings: Export format = "Comma separated file, .csv (e.g. Excel)" and Output = "Complete format". In result we got 10 CSV files of overall size ca. 77 MB.

Please note that although the above queries do not return disjoint bibliographic data, we will see that **CITAN** easily copes with such situation.

4.1. Importing data retrieved from *Scopus*

Let us now import the information on all the sources covered by *Scopus* and their ASJC codes. The whole list is available in the **CITAN** package (this data are in public domain — information from Elsevier).

```
> Scopus_ImportSources(conn); # import Scopus source coverage information
```

```
Importing Scopus ASJC codes... OK, 334 records added.
Importing country list... OK, 112 records added.
Importing Scopus source list... OK, 29913 of 30017 records added;
53762 ASJC codes processed.
```

A few of the sources (mainly conference proceedings) have not been added due to incorrect or lacking ISSNs.

We are now ready to import the CSV files to our LBS. As *Scopus* does not fully conform to the CSV standard, in case of a few records we will get some parse errors (each of them is indicated by the *Scopus_ReadCSV()* function). However, they may be easily corrected using a plain text editor, e.g. *Notepad*. Most of them concern improperly escaped single-quote characters.

```
> data <- Scopus_ReadCSV("Hirsch.csv"); # read the first CSV file (Query #1)
> lbsImportDocuments(conn, data, "ExampleSurvey") # import documents into LBS
Importing documents and their authors...
OK, 3 of 3 records added to ExampleSurvey/Hirsch.csv.
```

The first CSV file has been imported. Now, let us read the results returned by Query #2.

```
> data <- Scopus_ReadCSV("Hirsch_CITE.csv");
> lbsImportDocuments(conn, data, "ExampleSurvey")
Importing documents and their authors...
OK, 794 of 795 records added to ExampleSurvey/Hirsch_CITE.csv.
```

We see that one of the documents was not added because it was provided by the first file. However, an update process was triggered here. Such a behavior is especially useful when the data are downloaded in larger time spans and there is a possibility that the citation information has been changed. Let us proceed with the other CSV files.

```
> ... # read and import other CSV files
OK, 1933 of 1941 records added to ExampleSurvey/BibSciWeb_2002-2006.csv.
OK, 1353 of 1452 records added to ExampleSurvey/BibSciWeb_2007n_2011.csv.
OK, 1388 of 1574 records added to ExampleSurvey/BibSciWeb_2009n.csv.
OK, 1903 of 1903 records added to ExampleSurvey/BibSciWeb_BEf2002.csv.
OK, 1492 of 1804 records added to ExampleSurvey/5journals_1998-2005.csv.
OK, 291 of 353 records added to ExampleSurvey/5journals_2006.csv.
OK, 1355 of 1925 records added to ExampleSurvey/5journals_AFT2006.csv.
OK, 1555 of 1887 records added to ExampleSurvey/5journals_BEf1998.csv.
```

The whole import process takes about 21 minutes.

4.2. Cleaning duplicate entries

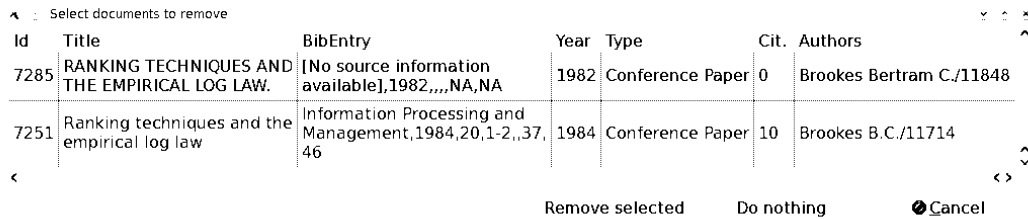
As it was mentioned above, **CITAN** uses *Scopus* document identifiers (*eids*) extracted from the *Link* field in the CSV files and author names to identify documents and authors, respectively. Unfortunately, the data exported from *Scopus* have sometimes ambiguous representation. For example, two versions of the same document may exist (e.g. an additional “in-press” copy) or Marie Salome Skłodowska-Curie may be stored in the LBS as Skłodowska M., Skłodowska-Curie M.S., or even Skłodowska M.S.C..

In order to detect such problems, we have implemented two functions which indicate potential groups of objects to be corrected. The results are presented in dialog boxes where we can easily choose the desired action on the selected records. Afterwards, we may apply the changes to the LBS using additional helper methods. Note that **CITAN** implements experimental versions of the disambiguation search algorithms; they currently trigger too many “nuisance alarms” and cause the process of data cleaning to be time-consuming. Their improvement is left to the further research. However, as we shall see on the screen-shots below, the results are promising.

Let us first take care of duplicated documents. The current version of the search algorithm makes use of a certain fuzzy similarity measure between the documents’ titles.

```
> ident <- lbsFindDuplicateTitles(conn, aggressiveness=2,
  ignoreTitles=c("In this issue%", "%In this issue", "%Editorial",
    "%Introduction", "Letter to %", "%Preface"));
```

Fig. 2 shows an exemplary search result presented by the above function. Please note the problem with the author’s name.



Id	Title	BibEntry	Year	Type	Cit.	Authors
7285	RANKING TECHNIQUES AND THE EMPIRICAL LOG LAW.	[No source information available],1982,,,,NA,NA	1982	Conference Paper	0	Brookes Bertram C./11848
7251	Ranking techniques and the empirical log law	Information Processing and Management,1984,20,1-2,,37,46	1984	Conference Paper	10	Brookes B.C./11714

Remove selected Do nothing ☒ Cancel

Figure 2: A group of duplicated documents indicated by *lbsFindDuplicateTitles()*.

We have marked 43 documents for removal. The following function applies appropriate changes to the LBS.

```
> lbsDeleteDocuments(conn, ident); dbCommit(conn);
```

We call a similar procedure for the authors.

```
> ident <- lbsFindDuplicateAuthors(conn, aggressiveness=2)
```

We have chosen 56 author groups to be merged. Two exemplary results are depicted in Fig. 3. To apply the changes, we then call:

```
> lbsMergeAuthors(conn, ident); dbCommit(conn);
```

Dialog 1: Select authors to merge

Parent	Child	Id	Name	List doc.	Documents	Citations
	✓	11620	Glaenzel W.		1	4
	✓	15601	Glaknzal W.		1	12
	✓	15830	Glanze W.		1	34
•		49	Glanzel W.		125	2342
	✓	11413	Glazel W.		1	4

Buttons: Merge selected, Do nothing, Cancel

Dialog 2: Select authors to merge

Parent	Child	Id	Name	List doc.	Documents	Citations
	✓	1083	Van Leeuwen T.		7	53
	✓	11418	Van Leeuwen T.H.N.		1	3
•		92	van Leeuwen T.N.		29	608
	✓	4406	Van Leeuwen Th.N.		5	160

Buttons: Merge selected, Do nothing, Cancel

Figure 3: Two groups of duplicated authors indicated by *lbsFindDuplicateAuthors()*.

4.3. Descriptive statistics

The *lbsDescriptiveStats()* function performs basic description of the sample. The graphical output is presented in Figs. 4–8.

```
> lbsDescriptiveStats(conn, surveyDescription="ExampleSurvey");
```

```
Number of sources in the database:    29913.
```

```
Number of documents in the database: 12024.
```

```
Number of authors in the database:    16282.
```

```
Data set restrictions:
```

```
Survey:          ExampleSurvey.
```

```
Document types:  <ALL>.
```

```
...
```

Fig. 4 presents a bar plot of document types in the LBS. To analyze authors' impact we will only use the following types: **ar** (**A**rticle; 8353 documents), **re** (**R**eview; 1110), **no** (**N**ote; 194), **sh** (**S**hort **S**urvey; 90). It is highly possible that they indicate high-quality peer-reviewed publications (9747 documents and 13953 authors overall).

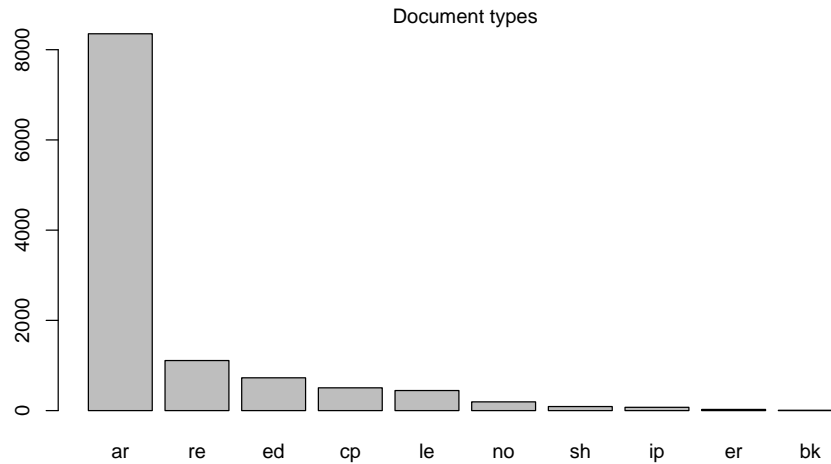


Figure 4: Descriptive statistics: Document types.

Assuming that all the documents published in the same journal or conference proceedings have a coherent subject-matter, we may try to classify them using the ASJC codes of their sources. Fig. 5 presents a pie chart of the upper-level categories. We see that the majority of the documents in our

sample are in computer science, the social sciences, medicine and engineering. There are also some publications in mathematics and decision support systems.

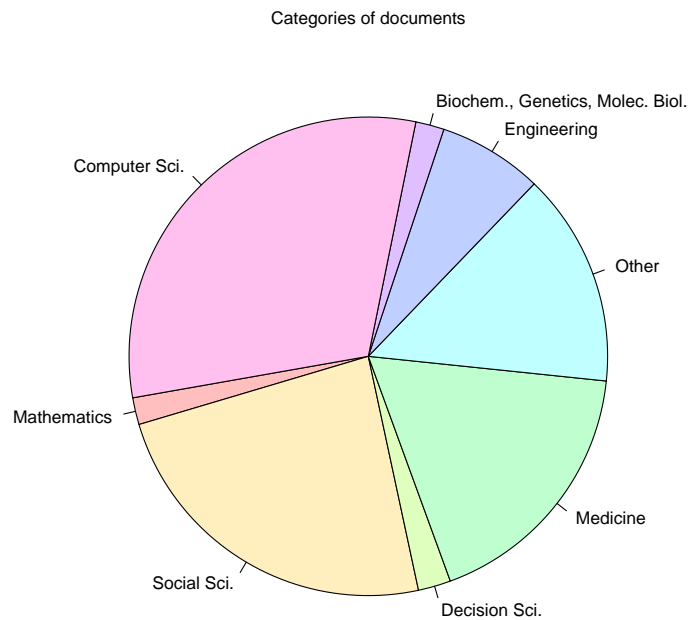


Figure 5: Descriptive statistics: ASJC document categories.

A box-and-whisker plot of page counts per document types is presented in Fig. 6.

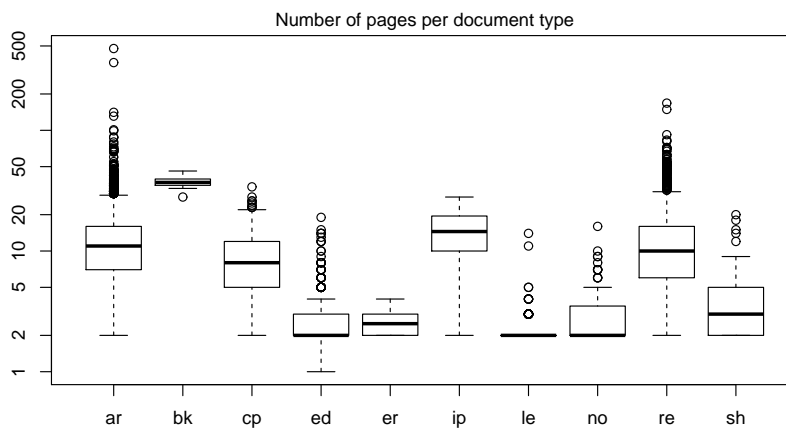


Figure 6: Descriptive statistics: Number of pages per document type.

Furthermore, Fig. 7 presents a box plot of citation counts per document type.

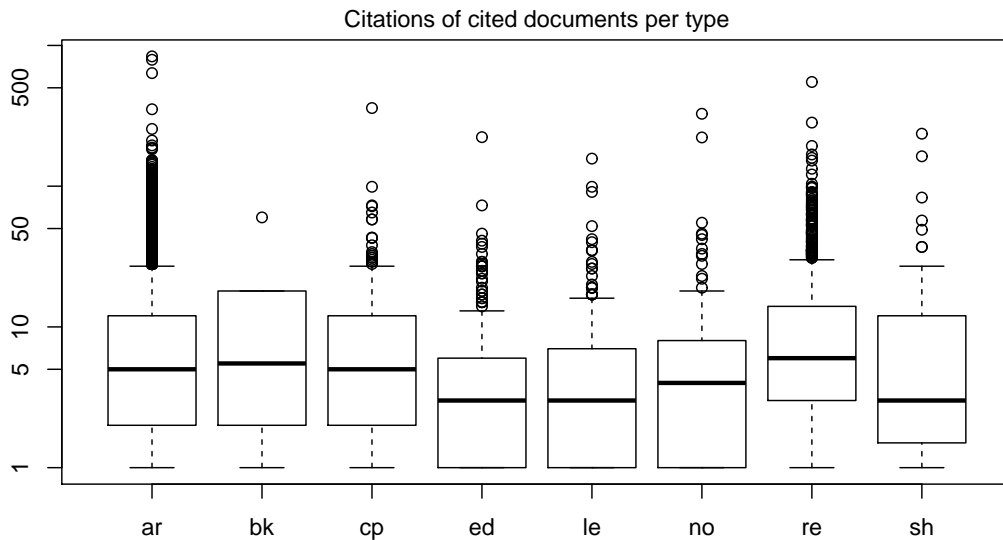


Figure 7: Descriptive statistics: Number of citations per document type.

Let us here verify a circulating fact (cf. Glänzel, 2008) that review papers are more often cited than ordinary articles.

First we extract citation data for all the documents of type **Article**.

```
> id_ar <- lbsSearchDocuments(conn, surveyDescription="ExampleSurvey",
  documentTypes="Article"); # search for appropriate documents
> info_ar <- lbsGetInfoDocuments(conn, id_ar); # get document information
> cit_ar <- sapply(info_ar, function(x) x$Citations); # extract citation counts
> summary(cit_ar); # descriptive statistics
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	3.000	8.716	9.000	836.000

The same should be done for **Reviews**.

```
> id_re <- lbsSearchDocuments(conn, surveyDescription="ExampleSurvey",
  documentTypes="Review");
> info_re <- lbsGetInfoDocuments(conn, id_re);
> cit_re <- sapply(info_re, function(x) x$Citations);
> summary(cit_re);
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	1.00	5.00	11.39	12.00	550.00

We shall use the non-parametric Wilcoxon rank sum test (as the distribution of citations is unknown) with a one-sided alternative hypothesis.

```
> wilcox.test(cit_ar, cit_re, alternative="less");
```

Wilcoxon rank sum test with continuity correction

```
data: cit_ar and cit_re
W = 4135932, p-value = 1.886e-09
alternative hypothesis: true location shift is less than 0
```

We see that the null hypothesis should be rejected at any sensible significance level ($W = 4135932$, $p\text{-value} \simeq 0$). Therefore we conclude that the articles are *generally* less often cited than review papers.

Finally, the two plots in Fig. 8 illustrate the number of documents as a function of publication year. We observe an exponential growth of the number of items in time, which is concordant to the results obtained by other researchers (cf. e.g. Gupta et al., 1995). A least squares fit for the time span 1973–2010 (using R call $lm(\log(y) \sim x)$) resulted in the model $y = \exp\{0.0969x - 187.8\}$ (a nonlinear model transformed to a linear case; determination coefficient $R^2 = 0.9456$).

Note that we may observe two interesting “jumps” on the histogram, located at ca. years 1996 and 2006. We could try to explain them by the expansion of the Scopus source coverage. However, by calculating an exponential least squares model for period 1996–2010 we obtain $y = \exp\{0.1137x - 221.4\}$ ($R^2 = 0.9629$), which fits better to the publication counts of the most recent papers.

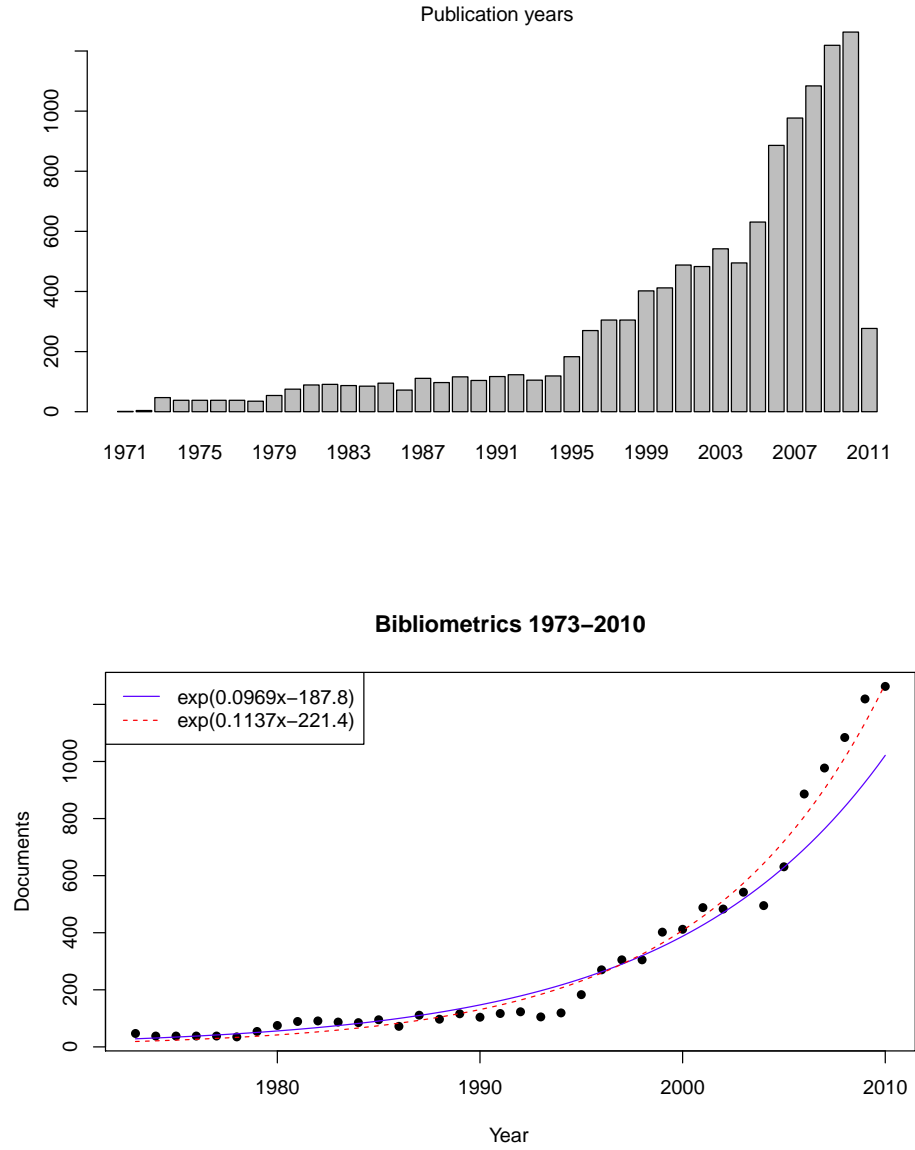


Figure 8: Descriptive statistics: Number of documents as a function of publication year (above). Below exponential least squares fits for period 1973–2010 ($y = \exp\{0.0969x - 187.8\}$) and for period 1996–2010 ($y = \exp\{0.1137x - 221.4\}$).

4.4. Assessing the impact of researchers

We are ready to perform the impact assessment of the authors in our LBS. As we have already mentioned, we restrict ourselves only to 4 document types which possibly indicate peer-reviewed publications.

First we generate the sequences $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$, where \mathbf{x}_i is a citation record of the i th author.

```
> citseq <- lbsGetCitations(conn, surveyDescription="DR",  
  documentTypes=c("Article", "Note", "Review", "Short Survey"));  
> length(citseq);    # calculate m, i.e. the number of authors
```

[1] 13953

The whole process takes about 3.5 minutes. The `citseq` variable is an object of class `list` consisting of numeric vectors. We see that not all of the 16282 authors in the survey were taken into account. This is because their publications did not match the *documentTypes* search criterion.

Let us determine the values of some impact functions for all the authors. We will use the following aggregation operators:

- (a) n (productivity),
- (b) \max (maximal citation count),
- (c) sum (overall citation count),
- (d) h (the h -index),
- (e) g (the g -index),
- (f) r_1 (an r_p -index; $p = 1$; equivalent to the w -index),
- (g) $l_1^{(g)}$ (an l_p -index projected to \mathbb{R} ; $p = 1$),
- (h) $l_\infty^{(g)}$ (an l_p -index projected to \mathbb{R} ; $p = \infty$),
- (j) $Ge1 = |\{x_j : x_j \geq 1\}|$ (a quasi-S-statistic, see Gagolewski and Grzegorzewski, 2011a),
- (k) $Ge5 = |\{x_j : x_j \geq 5\}|$ (a quasi-S-statistic),
- (l) $SLp1 = \sum_j \ln(x_j + 1)$ (a quasi-L-statistic).

Function call:

```
> res <- lbsAssess(
  citseq,          # a list of citation sequences
  bestRanks=19,    # limit results to top 19 values
  f=list(length, max, sum,          # (a-c)
    function(x) index.h(x),        # (d)
    function(x) index.g(x),        # (e)
    function(x) index.rp(x,1),      # (f)
    function(x) sqrt(prod(index.lp(x,1))), # (g)
    function(x) sqrt(prod(index.lp(x,Inf))), # (h)
    function(x) length(x[x>=1]),    # (j)
    function(x) length(x[x>=5]),    # (k)
    function(x) sum(log(x+1))),      # (l)
  captions=c("n", "max", "sum", "h", "g", "rp1",
    "l1", "lInf", "Ge1", "Ge5", "SLp1")
);
```

The calculation bases only on data stored in RAM, therefore it completes very quickly (in less than 10 secs.). We limited the output to top 19 results for each of the impact functions (this behavior is controlled via the *bestRanks* parameter). As a result, 42 authors' records were printed out (see Tab. 2).

The 5 best values for each impact function are typeset in bold. We observe that 5 of the authors are distinguished in such a manner by means of at least 6 criteria.

A few authors with small productivity (no more than 3 publications) were included in the ranking due to high values of the **max**, $l_1^{(g)}$, or $l_\infty^{(g)}$ functions. Interestingly, some of them appeared here because they co-authored a intensively-cited review paper (Gambhir S.S., Czernin J., Schwimmer J., Silverman D.H.S., Coleman R.E., Phelps M.E., A tabulated summary of the FDG PET literature, *Journal of Nuclear Medicine* **42**(5), 1–93, 2001 — 550 citations). Such a sensitivity for individual, highly-cited documents may be viewed as a drawback of these impact functions. On the other hand, only these functions were able to indicate the high impact of J.E. Hirsch on the field of scientometrics.

Table 2: Results of authors' assessment.

Autor	n	max	sum	h	g	r_1	$l_1^{(g)}$	$l_\infty^{(g)}$	$Ge1$	$Ge5$	$SLp1$
Bar-Ilan J.	36	64	624	14	24	27	29.7	16.2	34	24	83.7
Bordons M.	35	79	543	12	23	22	25.1	14.7	31	21	71.1
Bornmann L.	42	108	597	12	24	23	24.5	15.5	35	21	76.0
Braun T.	68	137	1223	19	32	36	37.2	19.9	65	51	158.5
Chen H.	45	133	739	14	26	25	30.0	18.2	36	25	88.9
Coleman R.E.	2	550	569	2	2	2	23.9	23.5	2	2	9.3
Cronin B.	28	94	568	12	23	22	25.6	16.1	26	22	67.2
Czernin J.	1	550	550	1	1	1	23.5	23.5	1	1	6.3
Daniel H.-D.	34	108	575	12	23	21	24.5	15.5	29	20	68.6
Egghe L.	92	195	1010	16	29	29	29.0	16.2	74	45	149.2
Gambhir S.S.	1	550	550	1	1	1	23.5	23.5	1	1	6.3
Garfield E.	24	327	914	10	24	18	29.0	21.7	22	15	58.0
Glanzel W.	105	137	2104	25	40	48	52.6	27.8	95	78	254.7
Hirsch J.E.	3	791	903	2	3	3	30.3	28.1	3	2	12.5
Jansen B.J.	15	352	796	10	15	14	27.9	19.4	14	12	43.7
King D.A.	1	283	283	1	1	1	16.8	16.8	1	1	5.6
Kostoff R.N.	52	147	840	16	27	32	32.0	17.3	47	38	115.8
Lewison G.	39	28	423	14	18	27	27.0	14.4	35	27	80.6
Leydesdorff L.	99	77	1382	21	32	42	42.2	22.3	89	69	211.9
Luwel M.	17	49	284	12	16	16	20.7	12.0	16	16	44.6
McCain K.W.	18	255	497	11	18	15	19.7	16.0	15	12	41.8
Meyer M.	23	62	435	12	20	21	24.2	14.7	21	19	57.2
Moed H.F.	49	133	1201	19	33	37	37.8	21.2	46	40	132.7
Moya-Anegon F.	38	40	262	10	14	17	17.9	10.5	32	16	59.0
Newman M.E.J.	1	836	836	1	1	1	28.9	28.9	1	1	6.7
Phelps M.E.	1	550	550	1	1	1	23.5	23.5	1	1	6.3
Rousseau R.	91	113	1193	19	31	34	34.6	19.8	80	49	171.4
Saracevic T.	11	352	825	10	11	11	28.6	18.8	11	11	42.1
Schubert A.	89	137	1563	22	36	42	42.5	22.5	79	62	197.8
Schwimmer J.	1	550	550	1	1	1	23.5	23.5	1	1	6.3
Seglen P.O.	9	636	836	6	9	9	27.1	25.2	9	9	28.7
Silverman D.H.S.	1	550	550	1	1	1	23.5	23.5	1	1	6.3
Spink A.	25	352	1076	18	25	24	36.2	20.3	24	24	81.4
Thelwall M.	86	96	1249	21	32	37	39.9	22.0	75	55	177.1
van Leeuwen T.N.	36	133	660	14	25	25	28.0	16.2	31	23	79.4
van Raan A.F.J.	66	148	1512	23	37	42	43.9	23.5	59	49	164.8
Vinkler P.	32	61	399	12	19	24	24.0	13.0	29	21	67.2
Wenneras C.	1	222	222	1	1	1	14.9	14.9	1	1	5.4
White H.D.	17	255	628	10	17	15	25.3	16.4	15	13	45.9
Wilson C.S.	32	45	287	9	15	17	18.7	11.0	29	17	57.7
Wold A.	1	222	222	1	1	1	14.9	14.9	1	1	5.4
Wolfram D.	19	352	479	5	19	9	20.2	18.8	14	6	28.3

5. Conclusions

ISO 9000 defines *quality* as a degree to which a set of inherent characteristics fulfills some well-precised requirements. To maintain a high level of output in any domain, specific quality control mechanisms have to be applied. In particular, the very complex case of the assessment of scientific achievements often involves a proper aggregation of the number of citations received by an author's publications.

As we currently do not have any other sensible way to objectively determine the significance of individual papers, we expect the citation analysis to be still receiving much interest in the future.

In this paper we presented a package for bibliometric impact assessment that is integrated into a statistical computing software. We discussed how the practitioners may use it in their own surveys. An important feature of **CITAN** is the support for processing and cleaning data retrieved from *Scopus*, a database with currently the broadest peer-reviewed source coverage available. Additionally, it may be used with the R environment in more complex analyses in the field of scientometrics.

Future work on the package will definitely concern the creation of a more sophisticated graphical user interface, which may allow the users with very little knowledge of R to perform the most elementary operations on the LBS, the improvement of the documents and authors disambiguation algorithms, and the support for other bibliometric databases.

Acknowledgments. The author would like to thank his PhD adviser P. Grzegorzewski and his student M. Dębski for stimulating discussion. Also, the author would like to express his gratitude to anonymous referees for their suggestions, and K. Fokow for proof-reading of the manuscript.

References

- Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., Herrera, F., 2010. *hg*-index: A new index to characterize the scientific output of researchers based on the *h*- and *g*-indices. *Scientometrics* 82 (2), 391–400.
- d'Angelo, C. A., Giuffrida, C., Abramo, G., 2011. A heuristic approach to author name disambiguation in bibliometric databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology* 62 (2), 257–269.

- Baneyx, A., 2008. *Publish or Perish* as citation metrics used to analyze scientific output in the humanities: International case studies in economics, geography, social sciences, philosophy, and history. *Archivum Immunologiae et Therapia Experimentalis* 56, 363–371.
- Beliakov, G., 2011. Fast Computation of Trimmed Means. *Journal of Statistical Software, Code Snippets* 39 (2), 1–6.
- Benjamini, Y., Hochberg, Y., 1995. Controlling False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* 57 (1), 289–300.
- Blum, M., Floyd, R. W., Pratt, V., Rives, R. L., Tarjan, R. E., 1973. Time bounds for selection. *Journal of Computer and System Sciences* 7 (4), 448–460.
- Crawley, M. J., 2007. *The R book*. Wiley.
- Egghe, L., 2006a. An improvement of the h -index: the g -index. *ISSI Newsletter* 2 (1), 8–9.
- Egghe, L., 2006b. Theory and practise of the g -index. *Scientometrics* 69 (1), 131–152.
- Franceschini, F., Maisano, D. A., 2009. The Hirsch index in manufacturing and quality engineering. *Quality and Reliability Engineering International* 25, 987–995.
- Franceschini, F., Maisano, D. A., 2011. Structured evaluation of the scientific output of academic research groups by recent h -based indicators. *Journal of Informetrics* 5, 64–74.
- Gagolewski, M., 2011. *CITAN: CITation ANalysis toolpack*. R package version 2011.05-1. <http://www.ibspan.waw.pl/~gagolews/CITAN/>
- Gagolewski, M., Dębski, M., Nowakiewicz, M., 2009. Efficient algorithms for computing “geometric” scientific impact indices. Tech. Rep. RB/1/2009, SRI PAS, Warsaw, Poland.
- Gagolewski, M., Grzegorzewski, P., 2009. A geometric approach to the construction of scientific impact indices. *Scientometrics* 81 (3), 617–634.

- Gagolewski, M., Grzegorzewski, P., 2010. Arity-monotonic extended aggregation operators. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer-Verlag, pp. 693–702.
- Gagolewski, M., Grzegorzewski, P., 2011a. Axiomatic characterizations of (quasi-) L-statistics and S-statistics and the Producer Assessment Problem. In: *Proc. Eusflat-LFA 2011*. In press.
- Gagolewski, M., Grzegorzewski, P., 2011b. Possibilistic analysis of arity-monotonic aggregation operators and its relation to bibliometric impact assessment of individuals. *International Journal of Approximate Reasoning*. In press. doi:10.1016/j.ijar.2011.01.010
- Glänzel, W., 2006. On the opportunities and limitations of the H-index. *Science Focus* 1 (1), 10–11.
- Glänzel, W., 2008. Seven myths in bibliometrics. About facts and fiction in quantitative science studies. *COLLNET Journal of Scientometrics and Information Management* 2 (1), 9–17.
- González-Pereira, B., Guerrero-Bote, V.P., Moya-Anegón, F., 2010. A new approach to the metric of journals scientific prestige: The SJR indicator. *Journal of Informetrics* 4 (3), 379–391.
- Grabisch, M., Pap, E., Marichal, J.-L., Mesiar, R., 2009. *Aggregation functions*. Cambridge.
- Graham, R. L., 1972. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters* 1, 132–133.
- Guns, R., Rousseau, R., 2009. Real and rational variants of the h -index and the g -index. *Journal of Informetrics* 3 (1), 64–71.
- Gupta, B. M., Sharma, L., Karisiddappa, C. R., 1995. Modelling the growth of papers in a scientific speciality. *Scientometrics* 33 (2), 187–201.
- Harzing, A.W.K., van der Wal, R., 2008. Google *Scholar* as a new source for citation analysis? *Ethics in Science and Environmental Politics* 8(1), 62–71.

- Hirsch, J. E., 2005. An index to quantify individual's scientific research output. *Proceedings of the National Academy of Sciences* 102 (46), 16569–16572.
- James, D. A., 2010. **RSQLite**: SQLite interface for R. R package version 0.9-4. <http://CRAN.R-project.org/package=RSQLite>
- Jin, B., Liang, L., Rousseau, R., Egghe, L., 2007. The R- and AR-indices: Complementing the h -index. *Chinese Science Bulletin* 52 (6), 855–863.
- Kosmulski, M., 2006. A new Hirsch-type index saves time and works equally well as the original h -index. *ISSI Newsletter* 2 (3), 4–6.
- Kosmulski, M., 2007. MAXPROD — A new index for assessment of the scientific output of an individual, and a comparison with the h -index. *Cybermetrics* 11 (1).
- Lawrence, M., Temple Lang, D., 2010. **RGtk2**: A graphical user interface toolkit for R. *Journal of Statistical Software* 37 (8), 1–52.
- Mallig, N., 2010. A relational database for bibliometric analysis. *Journal of Informetrics* 4 (4), 564–580.
- Meho, L. I., Rogers, Y., 2008. Citation counting, citation ranking, and h -index of human-computer interaction researchers: A comparison between Scopus and Web of Science. *Journal of the American Society for Information Science and Technology* 59 (11), 1711–1726.
- Meho, L. I., Sugimoto, C. R., 2009. Assessing the scholarly impact of information studies: A tale of two citation databases — Scopus and Web of Science. *Journal of the American Society for Information Science and Technology* 60 (12), 2499–2508.
- Moed, H.F., 2010. Measuring contextual citation impact of scientific journals. *Journal of Informetrics* 4 (3), 265–277.
- Norris, M., Oppenheim, C., 2010. Peer review and the h -index: Two studies. *Journal of Informetrics* 4, 221–232.
- Prathap, G., 2010. Is there a place for a mock h -index? *Scientometrics* 84, 153–165.

- R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>
- van Eck, N. J., Waltman, L., 2008. Generalizing the h - and g -indices. *Journal of Informetrics* 2 (4), 263–271.
- Vieira, E. S., Gomes, J. A., 2009. A comparison of Scopus and Web of Science for a typical university. *Scientometrics* 81 (2), 587–600.
- Woeginger, G. J., 2008a. An axiomatic analysis of Egghe’s g -index. *Journal of Informetrics* 2 (4), 364–368.
- Woeginger, G. J., 2008b. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences* 56 (2), 224–232.
- Woeginger, G. J., 2010. An algorithmic comparison of three scientific impact indices. *Acta Cybernetica* 19, 661–672.
- Wu, Q., 2010. The w -index: A measure to assess scientific impact by focusing on widely cited papers. *Journal of the American Society for Information Science and Technology* 61 (3), 609–614.
- Yu, H., Davis, M., Wilson, C. S., Cole, F. T. H., 2008. Object-oriented data modelling for informetric databases. *Journal of Informetrics* 2 (3), 240–251.

Please cite this paper as:

Gagolewski M., Bibliometric impact assessment with R and the CITAN package, *Journal of Informetrics*, 2011, doi:10.1016/j.joi.2011.06.006.