Journal of Informetrics 6 (2012) 505–512

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Stata commands for importing bibliometric data and processing author address information

Lutz Bornmann[a,*], Adam Ozimek[b]

[a] Max Planck Society, Administrative Headquarters, Hofgartenstr. 8, 80539 Munich, Germany
[b] Econsult Corporation, 1435 Walnut Street, Suite 300, Philadelphia, PA 19102, United States

ARTICLE INFO

Article history:
Received 15 December 2011
Received in revised form 30 March 2012
Accepted 3 April 2012

Keywords:
Stata
Bibliometric toolbox
Spatial scientometrics
Geocoding

ABSTRACT

Given the recent trend in bibliometrics and information science to use increasingly complex statistical methods, it is necessary to have powerful toolboxes to work with data from Web of Science (Thomson Reuters). We developed such a toolbox with four specific commands for the statistical software package Stata. These commands refer to (1) the import of downloads from Web of Science to Stata, (2) the preprocessing of address information from authors of publications in the downloaded set, (3) the geocoding of address information, and (4) the calculation of the minimum and maximum distance between several co-authors of a single paper. An advantage of developing commands for an established and comprehensive statistical software package (like Stata) is that a large number of further commands are available for the analysis of bibliometric data. We will describe some of these useful commands as well.

1. Introduction

In a recently published paper, Gagolewski (2011) introduced CITAN, the CITation ANalysis package for the R statistical computing environment, which provides bibliometricians and information scientists with software for use in the preprocessing, cleaning, and calculating of popular scientific impact indices by using SciVerse Scopus (Elsevier) data. Given the recent trend in bibliometrics and information science to use increasingly complex statistical methods, it is necessary to have powerful toolboxes which enable bibliometricians and information scientists to do this. In addition to R, there are other frequently used statistical software packages available (e.g., SPSS and Stata). For these other packages specific toolboxes for bibliometricians and information scientists are also necessary. We developed such a toolbox with four specific commands for Stata (StataCorp., 2011), which will be described in the present paper. Whereas Gagolewski (2011) focused on data from SciVerse Scopus, our tools (commands) are designed for Web of Science (WoS, Thomson Reuters) data. These commands refer to (1) the import of downloads from WoS to Stata, (2) the preprocessing of address information from authors of publications in the downloaded set, (3) the geocoding of address information, and (4) the calculation of the minimum and maximum distance between several co-authors of a single paper. An advantage of developing commands for an established and comprehensive statistical software package (like Stata) is that a large number of further commands are available for the analysis of bibliometric data. We will describe some of these useful commands as well.

To demonstrate the commands we use a data set of papers published from 1989 to 2009 in information science. The data set is used to demonstrate the commands rather than to present information science results. However, it is interesting to see

[*] Corresponding author.
E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), adam@econsult.com (A. Ozimek).

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.04.002

Table 1
Number of articles published in journals which are included in this study (absolute and relative frequencies).

Journal	Absolute frequencies	Relative frequencies
Journal of the American Society for Information Science and Technology	2148	30.3
Scientometrics	1947	27.5
Information Processing & Management	1217	17.2
Journal of Information Science	857	12.1
Journal of Documentation	504	7.1
Information Research	318	4.5
Journal of Informetrics	94	1.3
Total	7085	100

in this presentation of the commands how information science journals differ in their citation counts, and how the distance between co-authors developed over several years in this field.

2. Methods

2.1. Download and installation of the commands

Each of the tools can be installed within Stata using the standard command installation procedures. This is done by entering `findit` followed by the command name into the Stata command window. For instance, to install the `wosload.ado` command enter `findit wosload` and follow the on-screen installation instructions. Thus to install all of the commands the following should be entered:

```
findit wosload
findit wosaddress
findit geocode
findit groupdist
```

2.2. The data set used

All papers with the document type "Article" were first retrieved from the WoS database which had been published between 1989 and 2009. To cover the core journals of information science we included the same journals as used earlier by Leydesdorff and Persson (2010, p. 1623): (1) *Information Processing & Management* (INFORM PROCESS MANAG), (2) *Information Research* (INFORM RES), (3) *Journal of the American Society for Information, Science and Technology* (J AM SOC INF SCI TEC), (4) *Journal of Documentation* (J DOC), (5) *Journal of Informetrics* (J INFORMETR), (6) *Journal of Information Science* (J INF SCI), and (7) *Scientometrics*. Since the *Annual Review of Information Science and Technology* publishes almost exclusively reviews, we did not include this journal in the download. The search in WoS resulted in 7085 papers, which were saved in packages each containing 500 papers. Table 1 shows the number of papers per journal. As the table shows most of the papers were published in the *Journal of the American Society for Information Science and Technology* ($n = 2148$) and *Scientometrics* ($n = 1947$).

3. Presentation of the commands

3.1. Command wosload.ado

The basic command of the bibliometric toolbox is `wosload`. Downloads from WoS are saved as "Full record" (with or without Cited References) to "Tab-delimited (Win)"-files. Since no more than 500 records can be downloaded, more than one package (e.g., savedrecs1.txt, savedrecs2.txt, and savedrecs3.txt) must be saved. In this study, we downloaded 15 packages and saved them in one folder. With the command `wosload c:\p1 c:\p2 c:\p3 c:\p4 c:\p5 c:\p6 c:\p7 c:\p8 c:\p9 c:\p10 c:\p11 c:\p12 c:\p13 c:\p14 c:\p15` fifteen packages are imported from "c:\" into Stata and are combined to one data set (in Stata format), which can be saved as a whole. `wosload` requires the full path for each package and its filename without file extension (.txt). Other filenames than the default name used by Thomson Reuters (savedrecs) can be chosen. A variable `file` is generated which specifies the source package for each imported publication. Because `wosload` will import address fields in multiple variables, the address field (`c1`) has no limit on its length. However, all other fields are limited to Stata's standard 244 character length. This means that some variables that go beyond 244 characters, for example the abstract variable, will be truncated. `wosload` reports which string variables potentially are truncated, and for each of these variables it creates a dummy variable that indicates which observations may be long. The variable `c1` with the authors' addresses is not checked since it can be processed further using `wosaddress`. The dummy variables are named by appending "_long" to the end of the original variable name. So, e.g., if the variable with the author names (`au`) has some publications that are potentially truncated, a variable `au_long` is created. This variable will be equal to one for every publication that