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

Table 2
Papers in journals categorized by Thomson Reuter's journal classification scheme (absolute and relative frequencies).

Journal	Computer science, information systems		Computer science, interdisciplinary applications		Information science & library science	
	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.
Journal of the American Society for Information Science and Technology	2148	45.5			2148	30.3
Scientometrics			1947	100.0	1947	27.5
Information Processing & Management	1216	25.7			1216	17.2
Journal of Information Science	857	18.1			857	12.1
Journal of Documentation	504	10.7			504	7.1
Information Research					318	4.5
Journal of Informetrics					94	1,3
Total	4725	100.0	1947	100.0	7084	100.0

Note: one paper was not classified by Thomson Reuters.

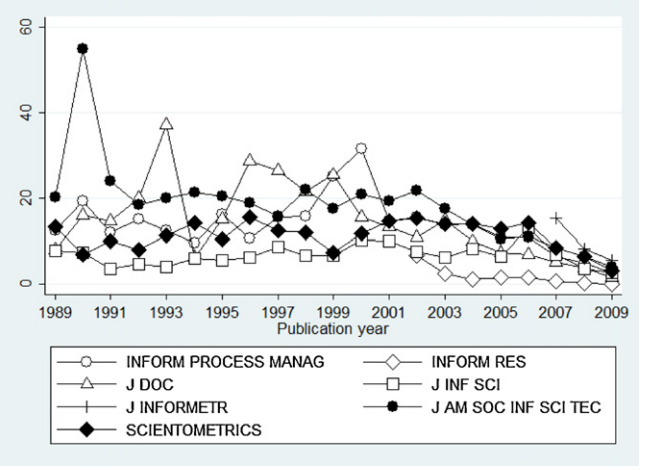

Fig. 1. Citation rates of papers published in different journals (arithmetic averages for papers published in one year).

is potentially truncated, and zero for those that are not truncated. It is important to note that these publications are only potentially truncated. If a publication by chance happens to be 243 or 244 characters long, it will be tagged as potentially truncated.

The variables in the imported data set are labelled according to the definitions of Thomson Reuters, which can be found here: http://images.webofknowledge.com/WOKRS53B4/help/WOK/hs_wos_fieldtags.html. The two-character field tags are used as variable names, and the field descriptions as variable labels. The labelling is optimized on publication sets containing papers with the document type "Articles" and "Conference Proceedings". In the following some Stata commands are described to further analyse the data.

There are some string variables (for example, `subject category` (`sc`) or `author keywords` (`de`)) in the data set downloaded from WoS with multiple parts containing different units of information that are separated by semicolons. For example "AGOSTI, M; GRADENIGO, G; MARCHETTI, PG" in the variable `authors` (`au`) represents author 1: AGOSTI, M; author 2: GRADENIGO, G; and author 3: MARCHETTI, PG. The command `split` can be used in Stata to split the contents of these string variables into more than one part. For example the variable `wc` (`Web of Science Category`) contains for each record one or more keywords that represent the publishing journal's classification (used by Thomson Reuters). Using the command `split wc, p(;)` with our data set, the content of `wc` is separated into two different variables (`wc1` and `wc2`). Since the categories in `wc2` start with a blank, they should be deleted by using the command `ltrim` (`replace wc2 = ltrim(wc2)`) to have consistent categories in both variables (`wc1` and `wc2`). The command `multencode` (here: `multencode wc1-wc2, gen(rwc1-rwc2) label(wc)`) creates new numeric variables (`rwc1` and `rwc2`), with value labels defined and attached that are based on the string variables (here: `wc1` and `wc2`). The same set of value labels is used for all new variables (`rwc1` and `rwc2`). Table 2 shows the number of papers (absolute and relative) in different information science journals (variable: `so`) categorized by Thomson Reuter's journal classification scheme. Since we have multiple categories which are stored in more than one variable (`rwc1` and `rwc2`) the command `mrtab` (Jann, 2005) is used to generate the table (`mrtab rwc1-rwc2, response(1/3) by(so) poly row`).

Fig. 1 shows mean citation rates for papers published in different publication years and journals. The figure has been generated by the `graph twoway scatter` command of Stata. Since the figure does not clearly show which journals' citation rates exhibit a statistically significant difference from one another, a regression model is calculated in a second step to answer

Table 3
Pairwise comparisons between the citation impact of different information science journals.

Journal	Citation rates (arithmetic average) across all publication years	. . . received statistically significantly **more** citations than:	. . . received statistically significantly **fewer** citations than:
Information Processing & Management (1)	12.4	2, 4, 7	6
Information Research (2)	1.0		1, 3, 4, 5, 6, 7
Journal of Documentation (3)	12.8	2, 4	6
Journal of Information Science (4)	6.7	2	1, 3, 6, 7
Journal of Informetrics (5)	9.7	2	6
Journal of the American Society for Information Science and Technology(6)	15.7	1, 2, 3, 4, 5, 7	
Scientometrics (7)	10.7	2, 4	1, 6

Note: significance level is $p < .05$ (Bonferroni adjusted).

this question. The outcome variable (here: number of citations) of this model is a count variable, which indicates "how many times something has happened" (Long & Freese, 2006, p. 350). The Poisson distribution is often used to model information on counts. However, this distribution rarely fits into the statistical analysis of bibliometric data, due to overdispersion. "That is, the [Poisson] model underfits the amount of dispersion in the outcome" (Long & Freese, 2006, p. 372). Since the standard model to account for overdispersion is the negative binomial (Hausman, Hall, & Griliches, 1984), we calculated negative binomial regression models in the present study.

The model (command `nbreg`) takes the number of citations (variable: `tc`) as dependent and the journals (variable: `so`) as independent variables (dummy variables). The publication years (variable: `py`) of the papers are included in the model predicting citation counts as exposure time (Long & Freese, 2006, pp. 370–372). We use the exposure option provided in Stata to take into account the time that a paper is available for citation. Pairwise comparisons using Bonferroni's adjustment for the *p*-values and confidence intervals are calculated to test which journals' citation impact exhibit a statistically significant difference. The exact commands for the analyses are as follows: `nbreg tc i.so, nolog exposure(py)` and `pwcompare so, effects mcompare(bonferroni)`. The results of the pairwise comparisons are presented in Table 3. Two journals exhibit a statistically significant difference from all other journals: the articles published in *Information Research* have been cited to a statistically significant lesser extent than the articles published in all other journals; for the articles published in the *Journal of the American Society for Information*, *Science and Technology* it is the other way round.

3.2. Command wosaddress.ado

Many bibliometric analyses use address data given by the authors on publications. For example, all comparisons of countries in terms of output and citation impact are based on the country information in the address field of WoS. The second command of the bibliometric toolbox, which we would like to introduce here is `wosaddress`. The programs `egenmore` and `renvars` must be installed in Stata in order to run this command. `wosaddress` converts data from the wide format to the long format so that each address (variable: `c1`) which is given on a single paper is on a separate line. Lines with different addresses for one paper are given a unique paper-based identification number in the new variable `id_wos`. The complete address of an author is stored in the variable `c1` (e.g., "Chinese Acad Sci, Grad Sch, Beijing, Peoples R China"); a short version of the complete address with only city and country information is in `address` (here: "Beijing, Peoples R China"). Besides these two address variables, three further variables are generated: (1) `address_count` contains the number of addresses on a given paper, `author_count` the number of authors, and (3) `country` the country information as one part of each address.

The running of `wosaddress` is a pre-condition for the running of two further commands which are included in the bibliometric toolbox and described in the following two chapters. However, the address information itself can be analysed by the command `screening`, which was developed by Belotti and Depalo (2010). This command examines "the content of complex narrative-text variables to identify one or more user-defined keywords" (Belotti & Depalo, 2010, p. 458). As an example for the benefit of `screening` in bibliometrics, we use the following command to identify addresses in `c1` from Amsterdam, Budapest, and Zurich: `screening, source(c1, upper) keys(zurich budapest amsterdam) explore(count) newcode(city, replace)`. The analysis is restricted to the publication years 1999–2009 since we have the impression that Thomson Reuters has not included the addresses of the reprint authors into the field `Author Address` (`c1`) in earlier years. We are interested in citation impact differences between papers published from authors located in the three cities. Fig. 2 shows the distributions of citations gathered from papers published by authors located in Amsterdam ($n = 102$ addresses), Budapest ($n = 110$ addresses), and Zurich ($n = 51$ addresses) (command: `graph box tc, over(city, descending relabel(1 "Zurich" 2 "Budapest" 3 "Amsterdam"))`). The horizontal line in the middle of each box indicates the median, and the top and bottom borders of the box mark the 75th and 25th percentiles, respectively. The whiskers above and below the box mark the upper and lower adjacent values. The points above the whiskers are defined as outliers. As Fig. 2 shows, the differences between the cities in terms of median citation rates are small (Amsterdam med = 10, Budapest med = 10.5, Zurich med = 12).

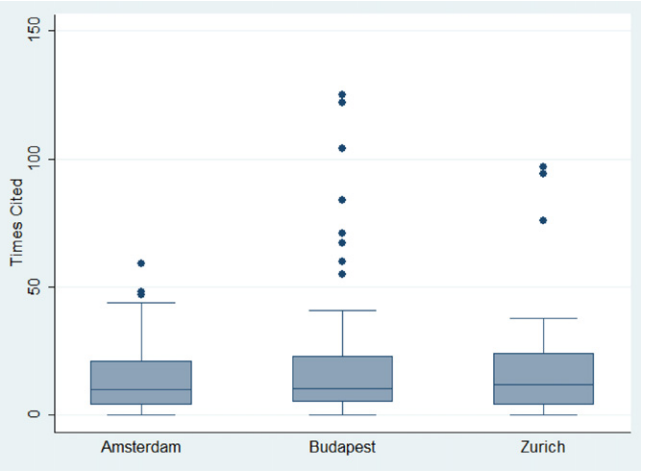

Fig. 2. Box plots for citations of papers which were published (between 1999 and 2009) by authors located in Amsterdam (n = 102 addresses), Budapest (n = 110 addresses), and Zurich (n = 51 addresses).

To confirm this result statistically, we calculate a regression model which is similar to the model described above. We use the following commands: `nbreg tc i.city, nolog exposure(py) cluster(id_wos)` and `pwcompare city, effects mcompare(bonferroni).city` is a variable that is generated by the above-mentioned `screening` command line containing the information as to whether an address belongs to Amsterdam, Budapest, or Zurich. In contrast to the other model, here we use `cluster(id_wos)` as an additional option. This option specifies that the addresses are independent across the papers, but are not necessarily independent within one and the same paper (Hosmer & Lemeshow, 2000). As the results of the pairwise comparisons following the model show, there are no statistically significant differences between the three cities in terms of citation impact.

3.3. Command geocode.ado

The command `geocode` automates the geocoding service included in Google Maps API and Yahoo! Maps API[1] to easily and quickly batch-geocode a set of addresses (here variable: `c1`). The command is an extended version of `geocode`, which was introduced by Ozimek and Miles (2011). The command `geocode, fulladdr(c1)` generates four new variables: `latitude`, `longitude`, `geocode`, and `geoscore`. `latitude` and `longitude` contain the geocoded coordinates for each address in decimal degrees, `geocode` contains a numerical indicator of geocoding success or type of failure, and `geoscore` provides a measure of the accuracy. The accuracy measure "indicates the *resolution* of the given result, but not necessarily the *correctness* of the result. For example, a geocode of '111 8th Avenue, New York, NY' may return '(Address) level accuracy,' indicating that the geocode is of the order of resolution of a street address. A geocode for 'France' would only return '(Country) level accuracy"' (http://code.google.com/intl/de-DE/apis/maps/documentation/geocoding/v2/#GeocodingAccuracy).

Table 4 shows the results of `geocode` and `geoscore` for the papers published between 1999 and 2009. The upper part of the table points out that geocodes could be retrieved for 87% of the addresses (n = 6389). The lower part shows that from the retrieved addresses approximately 96% are at least on the town accuracy level (accuracy level 4 or greater). The variables `latitude` and `longitude` in the data set can be used, for example, to visualize data as overlays on Google Maps (see here Bornmann & Leydesdorff, 2011; Bornmann, Leydesdorff, Walch-Solimena, & Ettl, 2011; Bornmann & Waltman, 2011). Focusing on only a part of Europe, Fig. 3 shows the spatially distributed addresses of authors who published papers in information science between 1999 and 2009. The whole map is based on 6389 addresses (see Table 4) and 1661 locations are unique (that means on average of 3.8 addresses per location). Different colours for the circles on the map indicate different numbers of addresses.

Since Yahoo! can also be used as a source for geocoding, the option `both` in the command `geocode, fulladdr(c1) both distm` specifies that coordinates are to be obtained not only from Google but also from Yahoo! as well (here: `ylat` and `ylon`) for a given data set. Similar to Google's geocoding procedure, Yahoo!'s quality scores are also added (`ygeocode` and `ygeoscore`). Information on what the two scores mean can be found here: http://developer.yahoo.com/geo/placefinder/guide/responses.html#address-quality. Google could not find geocoding coordinates for 989 addresses, while the corresponding number for Yahoo! was 226 (for 99 addresses both sources could not find any coordinates). To check the reliability of the geocoding results given by the two sources, the option `distm` in the

[1] Information on the Google Maps license can be found at https://developers.google.com/maps/terms and for the Yahoo Maps license at http://info.yahoo.com/legal/us/yahoo/maps/mapsapi/mapsapi-2141.html.

Table 4
Geocoding success or type of failure as well as geocoding accuracy for papers published between 1999 and 2009 (absolute and relative frequencies).

	Absolute frequencies	Relative frequencies
Google geocode definitions		
No errors	6389	86.6
Unknown address	926	12.6
No address specified	63	0.8
Total	7378	100.0
Google accuracy level for papers with "No errors"		
Country level accuracy (level 1)	34	0.5
Region (state, province, prefecture, etc.) level accuracy (level 2)	130	2.0
Sub-region (county, municipality, etc.) level accuracy (level 3)	91	1.4
Town (city, village) level accuracy (level 4)	2413	37.8
Post code (zip code) level accuracy (level 5)	3040	47.6
Street level accuracy (level 6)	37	0.6
Intersection level accuracy (level 7)	1	0.0
Address level accuracy (level 8)	7	0.1
Premises (building name, property name, shopping centre, etc.) level accuracy (level 9)	636	10.0
Total	6389	100.0

Fig. 3. Spatially distributed addresses of authors who have published papers in information science between 1999 and 2009 (*n* in the legend is the number of papers for a certain address). [Color figure can be viewed in the online issue].

Stata command specifies that the distances (in metres) between the Yahoo! and Google coordinates are calculated. Longer distances may indicate errors in geocoding. Table 5 shows median distances in kilometres between the geocoding results of Google and Yahoo! for addresses with different Google accuracy levels. The results indicate that the median distances for most of the accuracy levels are relatively low (e.g., levels 4 and 5). However, longer distances also occur, e.g., for accuracy levels 7 and 8 (with only 7 addresses). Accuracy level 2 seems to be the most problematic given the relatively high number of addresses ($n = 128$). It is interesting to see that the median distances do not correspond to the accuracy level: a higher level does not necessarily lead to shorter distances. Thus it seems sensible to use the comparison of Google and Yahoo! geocodes as a further quality check besides the quality scores provided by both.

Table 5
Median distances in kilometres between geocoding results of Google and Yahoo! for addresses with different Google accuracy levels (papers published between 1999 and 2009).

Google accuracy level	Absolute frequencies	Median
Country level accuracy (level 1)	34	10.35
Region (state, province, prefecture, etc.) level accuracy (level 2)	128	142.2
Sub-region (county, municipality, etc.) level accuracy (level 3)	91	0.61
Town (city, village) level accuracy (level 4)	2327	1.80
Post code (zip code) level accuracy (level 5)	3002	0.58
Street level accuracy (level 6)	37	5.12
Intersection level accuracy (level 7)	1	6403.6
Address level accuracy (level 8)	6	567.76
Premises (building name, property name, shopping centre, etc.) level accuracy (level 9)	36	1.51
Total	6262	

For $n = 127$ addresses Yahoo! could not find a corresponding geographic location.

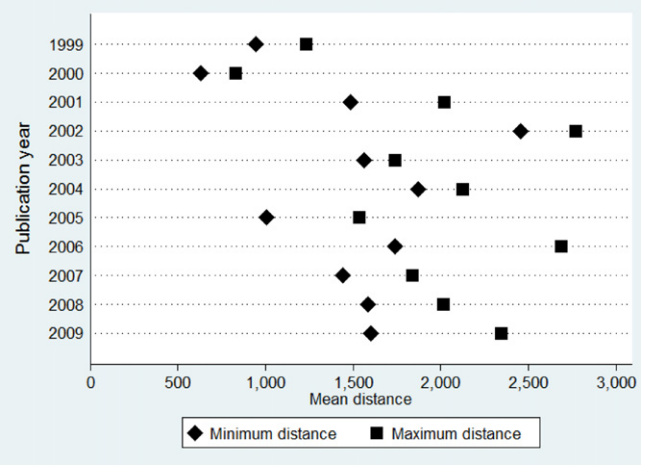

Fig. 4. Minimum and maximum distances (in kilometres) of single publication co-authors who published information science papers between 1999 and 2009. Each dot indicates an arithmetic average of distances for papers published within one year.

3.4. Command groupdist.ado

`groupdist` is the last Stata command we would like to introduce. It is intended to calculate the minimum or maximum distance between the authors of every single paper. The development of this command is motivated by a recent study of Waltman, Tijssen, and van Eck (2011). They measured "the extent and growth of scientific globalization in terms of physical distances between co-authoring researchers" (Waltman et al., 2011, p. 574). To demonstrate our command we restricted the dataset with papers published between 1999 and 2009 to those which have (1) at least two addresses, (2) all addresses geocoded, and (3) all addresses with a distance between the Google and Yahoo! geocodes of less than 30 km. These conditions restricted the data set to 1314 papers with 3070 addresses in total. We used the following command to calculate the minimum and maximum distance: `groupdist, x(longitude) y(latitude) group(id_wos) inkm`. To run the command in Stata, the program `vincenty` must be installed. The option `inkm` specifies that the distances are calculated in kilometres; otherwise, by default, the calculations are based on miles. The variable `id_wos` contains the unique identification number for the different addresses of one paper. The largest geographical distance (here referred to as maximum distance) between two addresses in a paper's address list is termed the geographical collaboration distance (GCD) by Waltman et al. (2011).

Fig. 4 shows minimum and maximum distances (in kilometres) of a single publication's co-authors for published information science papers between 1999 and 2009. We first determined the minimum and maximum distances between authors of every single publication and then calculated the mean minimum and maximum distances per year. Thus, each dot indicates an arithmetic average of distances for papers published within one year. The results do not indicate clear trends of increasing or decreasing minimum or maximum distances in information science. The correlations (product–moment correlation coefficients) between publication year and minimum distance ($r = .04$) as well as between publication year and maximum distance ($r = .07$) are very low. Thus, our results are in disagreement with the results of Waltman et al. (2011). They found that "science has globalized at a fairly steady rate. The MGCD [mean geographical collaboration distance] for science as a whole has increased more or less linearly over the past three decades from 334 km in 1980 to 1553 km in 2009" (Waltman

et al., 2011, p. 576). Whereas our analyses are based only on 1314 papers with 3070 addresses in total, the study of Waltman et al. (2011) included more than 20 million publications with just under 39.0 million addresses.

However, the study of Hennemann, Rybski, and Liefner (2012) shows that a notion of globalized scientific collaboration is not supported by their empirical data. Using a novel approach of analysing distance-dependent probabilities of collaboration, their "analysis of six distinct scientific fields reveal that intra-country collaboration is about 10–50 times more likely to occur than international collaboration" (p. 217).

4. Discussion

Following the demand for bibliometricians and information scientists to use more and more complex statistical methods, we introduce four Stata commands here (1) which can be used to easily import publication sets from WoS to Stata, (2) to preprocess address information given on publications for further processing, (3) to geocode author addresses, and (4) to calculate the minimum and maximum distances between several co-authors of a publication. With this paper, we follow activities like those of Gagolewski (2011) who introduced CITAN to the software R. For the future it is planned to extend the four commands in our toolbox with further options and to introduce further commands. For example, a command for the calculation of the *h* index and its variants (Bornmann, Mutz, Hug, & Daniel, 2011) may be of interest for bibliometricians and information scientists. We think it would be also interesting for the users to choose between full or fractionate counting of publications in `wosload`.

We highly appreciate feedback from users of our toolbox, and are interested in problems in running the commands and further options which could optimize the functionality. Ideas for further commands are also welcome.

References

Belotti, F., & Depalo, D. (2010). Translation from narrative text to standard codes variables with Stata. *Stata Journal*, *10*(3), 458–481.

Bornmann, L., & Leydesdorff, L. (2011). Which cities produce more excellent papers than can be expected? A new mapping approach—using Google Maps—based on statistical significance testing. *Journal of the American Society of Information Science and Technology*, *62*(10), 1954–1962.

Bornmann, L., Leydesdorff, L., Walch-Solimena, C., & Ettl, C. (2011). Mapping excellence in the geography of science: an approach based on Scopus data. *Journal of Informetrics*, *5*(4), 537–546.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H. D. (2011). A meta-analysis of studies reporting correlations between the *h* index and 37 different *h* index variants. *Journal of Informetrics*, *5*(3), 346–359. http://dx.doi.org/10.1016/j.joi.2011.01.006

Bornmann, L., & Waltman, L. (2011). The detection of hot regions in the geography of science: a visualization approach by using density maps. *Journal of Informetrics*, *5*(4), 547–553.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692. http://dx.doi.org/10.1016/j.joi.2011.06.006

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents R and D relationship. *Econometrica*, *52*(4), 909–938.

Hennemann, S., Rybski, D., & Liefner, I. (2012). The myth of global science collaboration—collaboration patterns in epistemic communities. *Journal of Informetrics*, *6*(2), 217–225. http://dx.doi.org/10.1016/j.joi.2011.12.002

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Chichester, UK: John Wiley & Sons, Inc.

Jann, B. (2005). Tabulation of multiple response. *The Stata Journal*, *5*(1), 92–122.

Leydesdorff, L., & Persson, O. (2010). Mapping the geography of science: distribution patterns and networks of relations among cities and institutes. *Journal of the American Society for Information Science and Technology*, *61*(8), 1622–1634. http://dx.doi.org/10.1002/Asi.21347

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX, USA: Stata Press, Stata Corporation.

Ozimek, A., & Miles, D. (2011). Stata utilities for geocoding and generating travel time and travel distance information. *Stata Journal*, *11*(1), 106–119.

StataCorp. (2011). *Stata statistical software: release 12*. College Station, TX, USA: Stata Corporation.

Waltman, L., Tijssen, R. J. W., & van Eck, N. J. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574–582. http://dx.doi.org/10.1016/j.joi.2011.05.003