ISSI Newsletter

QUATERLY E-NEWSLETTER OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

ISSN 1998-5460

#27 / VOLUME 07 NUMBER 3
SEPTEMBER 2011

CONTENTS

EDITORIAL

Dear members,

In the previous issue of the newsletter you could read the results of the elections for president and for member of the board. I welcome Leo Egghe, Henk Moed and Ed Noyons as new (or at least newly elected) members, and I thank Judit Bar-Ilan, Martin Meyer and Olle Persson for their contributions to the board (expressing the hope that they will continue contributing time and energy to ISSI). Aparna Basu, Peter Ingwersen and Grant Lewison are staying "on board" for another two years (at least, as they can be re-elected). I thank you all for your vote of confidence, electing me as your president for the second time. Be assured, there will not come a third time. Firstly, because I think that leadership of a society (a country, a journal, a company) should be renewed after some years, and secondly because the new board has followed me in this and decided that a president of our society can be re-elected at most once.

The main event of the preceding months was of course our biennial conference, held this time in Durban, South Africa. In this issue you may read some impressions and see some pictures taken during the conference. Many talks have started in Durban and surely they have continued and will continue further over the following months. In name of the society I thank the organizers for all their work, but referring to an African saying "it takes (the equivalent of) a whole village" to organize a conference and do all the things that must be done to assure that the whole event runs smoothly. So my thanks goes to this whole village of behind-the-scene collaborators.

In this issue you further find contributions by Jonathan Levitt about funding and one by Leo Egghe about Benford's law and the failure of the European Union to take this regularity into account. Our Dutch colleagues Nees Jan van Eck and Ludo Waltman present the latest version of VOSviewer and describe how their software can be applied in the context of text mining. Finally, Dilruba Mahbuba and me wrote an article about Bangladesh and its centre of publication.

I hope these contributions will provide inspiring reading to members and (in a few months) non-members alike.

Ronald Rousseau

ISSI e-Newsletter (ISSN 1998-5460) is published by ISSI (http://www.issi-society.info/).
Contributors to the newsletter should contact the editorial board by e-mail.

THE ISSI CONFERENCE IN SOUTH AFRICA

A REPORT ON THE 13TH ISSI CONFERENCE DURBAN (SOUTH AFRICA)

DENNIS OCHOLLA

PETER INGWERSEN

INTRODUCTION

South Africa successfully hosted the 13th ISSI International Conference, from the 4th to the 7th of June 2011 (http://www.issi2011.uzulu.ac.za), in the culturally-rich South African city of Durban. This was the first time the conference was held on African soil. We believe that a conference of this magnitude always encourages the growth of research and the increase of national and international collaboration and that we were able to popularize research in the areas of Informetrics, Scientometrics and webometrics within the country and on the continent. The ISSI 2011 Conference provides an international open forum for scientists, research managers and authorities, and information and communication related professionals to debate the current status and advancements of Informetrics and Scientometrics theory and applications, with emphasis on the progress of Scientometrics and science in developing countries. The conference was organized under the auspices of ISSI.

Broadly, although the term 'informetrics' has become increasingly popular, Scientometrics, Informetrics, bibliometrics and webometrics are all closely related or interlinked. This is due to the fact that they can all be employed for quantitative analysis or measurement of all forms of recorded information in pure, applied and action research. This can be achieved by studying their distribution, circulation and use pattern largely within (or among) individuals, disciplines, organizations or countries, for informing research, political, economic, scientific and technological issues, and knowledge policies and decisions (to name but a few). The Informetrics disciplines thus contribute to evidence-based and informed knowledge about scientific

© Photo copyright: Zsuzsanna Glänzel

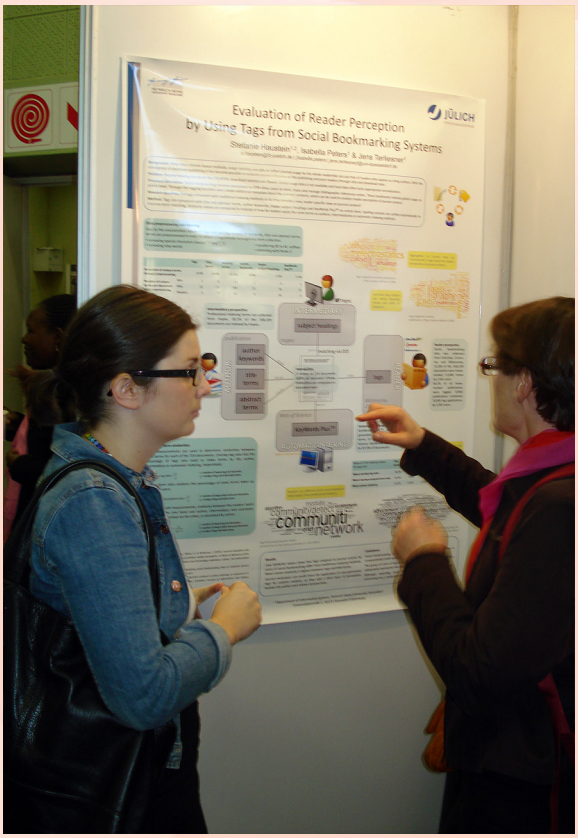
© Photo copyright: Zsuzsanna Glänzel

© Photo copyright: Zsuzsanna Glänzel

research and provide input to research and innovative policy-making worldwide.

Historically, since 1940, Informetrics has flourished and become a major sub-discipline within information science. There is a strong overlap between Informetrics, bibliometrics, Scientometrics, cybermetrics and webometrics, as illustrated by Bjorneborn and Ingwersen (2004).The role of the International Society for Scientometrics (ISSI), as it is widely understood, is to spearhead research, scholarly exchange and dissemination of bibliometrics, scientometrics, webometrics and informetrics information. The "Time Line of Bibliometrics" provided by Rousseau (2002), Hertzel (1987), Hood and Wilson (2001), as well as ISI (see http://vmw.isinet.com/isi/oboutl timeime.html) shows how far this discipline has grown with some significant details. A big part of publications, in this domain, are largely published in Scientometrics and the Journal of Informetrics.

ISSI regularly meet bi-annually. Previous ISSI conferences took place in Belgium (1987), Canada (1989), India (1991), Germany (1993), USA (1995), Israel (1997), Mexico (1999), Australia (2001), China (2003), Sweden (2005), Spain (2007) and Brazil (2009). The 13th ISSI conference – in South Africa – was largely organized by the University of Zululand, Durban University of Technology and the University of South Africa. Support was also received from other institutions such as the National Research Foundation (SA), University of Western Cape, University of Cape Town, University of Pretoria, University of KwaZulu Natal, Library and Information Association of South Africa (LIASA) and under the auspices of the International Society for Scientometrics and Informetrics (ISSI). The conference attracted 200 participants from 32 countries. In total, from 32 different countries, we received 175 submissions for Full papers and Research in Progress (RiP) papers and around 40 submissions for posters. At the conference, 65 full papers and 28 RiP papers were orally presented, together with 52 posters, also including revised full and RIP papers after peer review. Overall around 30% were rejected or withdrawn. The conference was



© Photo copyright: Jacqueline Leta

© Photo copyright: Jacqueline Leta

opened by the Vice Chancellor of the Durban University of Technology Prof. Ahmed Bawa. The conference keynote speakers were: Prof. Anastassios Pouris from South Africa, Prof. Olle Persson from Sweden (he was also the recipient of Derek de Solla Price Medal), Prof. Ricardo Baeza-Yates from Spain and Prof. Jonathan Adams from the United Kingdom.

CONFERENCE ORGANIZATION

The 13th ISSI conference, in South Africa, was organized by the ISSI 2011 Conference Local Organizing Committee, with the participation of six Universities and the National Research Foundation(NRF).These were : Durban University of Technology (the conference venue – http://www.dut.ac.za/site/default.asp), University of Cape Town (http://www.uct.ac.za), University of KwaZulu Natal (http://www.ukzn.ac.za), National Research Foundation (http://www.nrf.ac.za), University of South Africa (http://www.unisa.ac.za), University of Pretoria (http://www.up.ac.za) and University of Zululand [main organizer-(http://www.uzulu.ac.za)].

The conference Chairs were: Dennis N. Ocholla, Daisy Jacobs and Peter Ingwersen. Programme Chairs were: Ed Noyons and Patrick Ngulube; Poster Chairs: Jacqueline Leta and Bosire Onyancha and the Doctoral Forum Chairs were: Jesper W. Schneider and Bosire Onyancha. The Thematic chairs included: Theory – Liming Liang ; Methods and techniques – Ronald Rousseau ; Citation and co-citation analysis – Birger Larsen ; Indicators – Henk Moed ; Webometrics - Mike Thelwall; Mapping & visualization - Katy Börner ; Research policy – Éric Archambault; Productivity & publications – Wolfgang Glänzel ; Journals, databases and electronic publications - Bluma Peritz ; Collaboration – Aparna Basu ; Country level studies – Jane Russell and Patent analysis – Martin Meyer.

DURBAN AS VENUE

The choice of Durban, to be the conference city, was significant. Durban (http://www.

world66.com/africa/southafrica/durban/lib/gallery) is considered to be Africa's leading conference destination; is a vibrant city with a harmonious blend of African, Asian and European culture -often reflected in its tasteful cuisine - and is located in the historical Kingdom of the Zulu nation, in the KwaZulu-Natal province, that is a gateway to African culture in South Africa. Durban is also South Africa's only destination of tropical summers, with 320 sunny days a year. The Durban University of Technology, the conference venue that is located close to the city's Central Business District (CBD), is one of the 23 public universities in South Africa that focuses on technology oriented vocational and professional higher education, with 23,000 student's enrolment. The University allocated its conference facility for the ISSI conference.

Participation in the conference's social events was one of the most important events during the conference. The conference reception, organized at the Docklands Hotel Waterfront in Durban on July 4th and sponsored by Thomson-Reuters, highlighted the Marimba Music performance that was exhilarating and extremely entertaining as it brought even the most serious conference participants to the dance floor. The Conference Dinner at MOYO uShaka Restaurant (at uShaka Village park) was another memorable social event which combined African cuisine, dance, music and an extraordinary ocean environment (as the restaurant is located at Durban's beachfront) in the midst of an extremely welcoming and jovial restaurant staff. Many conference participants liked the conference venue at the Durban University of Technology (DUT) particularly the rich, diversified and generous food/lunches that were served/offered by the DUT Hotel School. One conference day-catering was sponsored by Thomson-Reuters. The visit, by some of the conference participants, to South African game reserves and other exciting tourist destinations – we were told – left unforgettable and memorable experiences that made this conference unique and even more successful.

THE CONFERENCE PROGRAMME

The conference papers stemmed from the following twelve sub-themes: Theory, methods and techniques, citation and co-citation analysis,indicators,webometrics,mapping & visualization, research policy, productivity & publications, journals, databases and electronic publications,collaboration,country level studies and patent analysis.

The pre-conference – held on the 4th of July – was dedicated to a Doctoral Forum that used experts in the discipline to discuss, help and advise doctoral students (of scientometrics and informetrics), selected from all over the world in the field, on how to develop their research proposals as well as workshops. Ten students participated in the forum. The two ISSI workshops/tutorials were, in total attended by 80 participants, focused on "Introduction to Scientometrics and Webometrics" by Peter Ingwersen and "Sci2: A Tool of Science of Science Research and Practice" by Dr. Katy Börner. The period from the 5th -7th of July was dedicated to the main conference (on the highlighted themes).

CONCLUSION

Aside from all these arrangements, we had an enthusiastic and able team of organizers and supporters (both here in the province, of KZN, and from other institutions in the country). The message we continue getting from participants is that 13th ISSI Conference in Durban was a big success more than was expected. We wish to acknowledge the support we received from the University of Zululand (in terms of staff time, transport and administrative costs); the National Research Foundation (NRF) (for partial funding for administrative costs); Durban University of Technology (for support with the confer-

© Photo copyright: Jacqueline Leta

ence venue and its administration) and the University of South Africa (for staff support). We would also like to convey gratitude to our sponsors: National Research Foundation (SA) for sponsoring administrative costs, Thomson Reuters (Conference reception and one day of the conference), Elsevier (part of conference proceedings) and other organizations who gave us small donations, such as SABINET, Yahoo Research, and many more. When we receive the outstanding sponsorships funds from some of the listed organizations, we will also balance our conference books that will be another dimension of success.

REFERENCES

Bjorneborn and Ingwersen. 2004. Towards a basic framework of webometrics. *Journal of the American Society for information Science and Technology*, 55(14), 1216-1227

Hertzel . 1987. Bibliometrics. History of the Development of ideas in statistical Bibliography or Bibliometrics. In: Kent, A &Lancour, H.(Eds), *Encyclopedia of Library and Information Science* Vol. 42, 144-219

Hood, W.W. and Wilson, C.S. 2001. The literature of bibliometrics, scientometrics and informetrics. *Scientometrics*, 52(2), 291 - 314

Rousseau, R. 2002. Timeline of Bibliometrics. [Online] http:// users.pandora.be/ronald.rousseau/htm/timetine of bibliometrics.html Accessed 15 July 2007.

© Photo copyright: Jacqueline Leta

INTRODUCING THE DEREK DE SOLLA PRICE AWARDEE OF 2011

INTERVIEW BY BALÁZS SCHLEMMER

The awarding ceremony of the Derek de Solla Price Memorial Medal has become an essential part of the programme of ISSI conferences since the foundation of the Society in 1993. The Price Medal was conceived and launched by Tibor Braun, founder and Editor-in-Chief of the international journal Scientometrics, and is periodically awarded by the journal to scientists with outstanding contributions to the fields of quantitative studies of science. This year's awardee is OLLE PERSSON (Inforsk, Department of Sociology, Umeå University). Congratulations to the award-winner!*

OLLE PERSSON

■ **Attentive readers may remember that interviews with earlier Price Awardees (see ISSI e-Newsletter #2, #11 and #17) usually began with a question targeting the awardees' crooked routes leading from their original majors or areas of research (biology, literature, chemistry etc.) to scientometrics – like if it should be taken for granted that a scientist could end up at bibliometrics only after (or parallel to) years of meandering on other fields of science. But it's not the case with you: your publication list suggests that you started your career directly with information science, more exactly, bibliometrics. How come?**

Let me start by summarizing my life. I was born in Luleå in 1949 and moved to Umeå in 1968. Since 1972 I've been at the same department (sociology), studying the same field (information science), married to the same woman (Ann-Britt) and lived on the same street (Axtorpsvägen, less than five minutes walk to the university). Believe me or not, I still enjoy every day! My bosses and my wife have always allowed me to do what I liked. But, I must admit, they have had some problems understanding what my research is about. However, the growing success of bibliometrics, and lately the Price award have made them more curious.

* You can learn more about the award and award winners on the ISSI website: http://www.issi-society.info/price.html

As a young doctoral student in sociology, in 1975, I received a special 3-year scholarship in Sweden for doing information science research. There was a need for a sociologist to study and evaluate the use of information retrieval systems. I was happy to get it, but my fellow sociology students really wondered what it was all about. The concept of information was a suspicious one they thought. Capital and labor were the cornerstones of society – but information, what is that? Now they know better.

■ **Do you still remember what the main findings of your first publications were?**

My first published paper in bibliometrics appeared in 1979. It was a study of the citations to, or rather mentions of, classical sociologist (Marx, Weber and Durkheim) in Sociological Abstracts. This paper is my first entry in Web of Science entitled "Classic sociological works – report from data center", however in Swedish. Using a tty-terminal (see picture) I logged on to the Dialog-system over the telephone network, and issued a series of search commands to get time series from 1963 for each of the three scholars. The result was amazing, showing the rise and fall of Marx, strongly reflecting the overall climate in social science of those days. Online bibliometrics was born, and later on I wrote a number of articles, chapters and reports on how to generate interesting data from on-line data base searches.

ASR-33 Teletype.
© Photo copyright: Courtesy of David Gesswein, http://www.pdp8online.com/

■ **What do you consider your most important publication? Not necessarily the one with the highest citation impact, but the one, which is your personal favorite just because of the complexity and/or beauty of the research.**

I think that the 1994 paper "The intellectual base and research fronts of JASIS 1986-1990" is my personal favorite. It was very exciting to develop the mapping tools for analyzing the knowledge base and research fronts of our discipline. Again I was struck by the strong face validity of the results. A few years later I was very happy when White & McCain in their 1998 paper cited me and wrote that I described "the same elephant" as they did.

■ **Have you ever had a very surprising research result which was completely against your preliminary expectations?**

Recently I published a paper "Are highly cited papers more international?" We have always heard that multi-country papers are significantly more cited than domestic papers. However, this study rather suggests that most of the highly cited papers in hot research areas are domestic. Working alone or in small tight local groups is still of great importance for breakthroughs in science.

■ **You were one of the founding members of Inforsk (The Information Research Group) in 1975, almost four decades ago. It must have been a very progressive and nearly only-one-of-its-kind institute that time. How did it all start and why? How did the work go? Tell us, please, about those heroic times with very limited IT infrastructure.**

At the Durban Conference in 2011.
© Photo copyright: Ann-Britt Persson

In the mid 1970's the Swedish government had the idea the access and use of scientific information was an important issue, and decided to support a few young researchers. The Inforsk group, primarily Lars Höglund and I, made a number of studies of information needs and uses in different organizations and subfields. I made a set of sociometric studies to show the importance of informal communication. Among other things it became quite obvious that the most read authors were also at the center of discussion networks. To reach the top you must climb both the bibliometric and the sociometric ladder.

During those early years we also collaborated with Thomas J Allen (MIT) on the informal networks of Swedish engineers in the field of iron and steel. Although we never co-published an article, Tom was very inspiring in helping us to replicate some of his studies on Swedish engineers. A few years later Tom, Maurice B Line and Osmo Wiio evaluated the Inforsk group and came out with a very positive review. I especially recall the positive words from Maurice on my early work in bibliometrics.

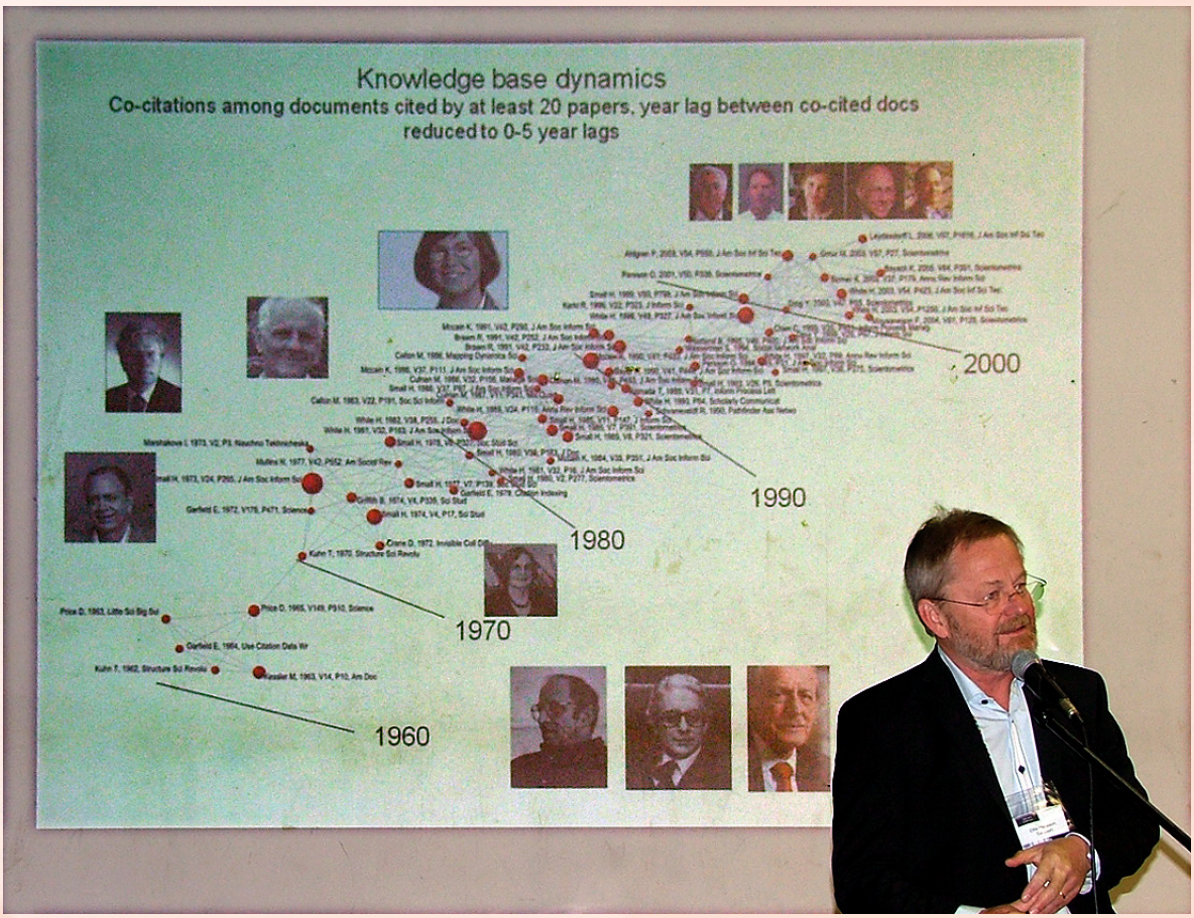
Up to mid 1980's we used terminals connected to main frame computers to analyze our data. My first PC was an Osborne, with two diskette stations, one for the CPM operating system that had to be read first, and then replaced by the Word-star diskette. The other station had a diskette for my data. The next one was an IBM XT that had a memory that could hold the data and programs even after it was turned off. What a giant step! Before the Internet we had also e-mail. However, we needed to make a special login to check for messages, and most of the time the box was empty

■ **You are widely known about your bibliometric tool, BibExcel, which is a free software for bibliometric analysis. When did you start to develop it and what gave you the idea? Do you have any feedback (e.g. number of downloads) about how many researchers use it worldwide?**

If you wish to make bibliometric research you need a tool box that allows you to turn bibliographic data into variables and numbers. Using on-line search systems do give you a chance to make some interesting observation, but to make more you need to work on downloaded records. Around year 2000 I started to develop BibExcel for my own needs. When I put BibExcel on the net I got lots of suggestions for improvements from collaborators and students. Now BibExcel has a lot of different tools that enables a great variety of analysis. For mapping purposes it also generates files for further processing in Pajek, Usinet, Mapequation etc. On the Inforsk homepage there is a link to "Cluster maps" that shows the visitors to the site, and I guess that they often download BibExcel from there.

I recently made a citation search of BibExcel using the Google Scholar interface called "Harzing's Publish of Perish" which yielded H-Index = 20, which is in fact greater than searching my own name.

Potential downloaders of BibExcel from all over the world



Olle in good company. Presentation on the occasion of being awarded by the Price Medal. ISSI 2011 Conference, Durban, South Africa. © Photo copyright: Ann-Britt Persson

■ **Now let's take a closer look at the man behind science: How do you think your colleagues and/or students characterize you? And how do you refine the picture?**
I hope that they think that I am a happy guy that likes to mix serious research with having fun. They might even think I am the altruistic type by offering my software and advice for free. But be in no doubt, I get a lot back from them!

■ **You are also very well-known from conferences and other similar events: you are the founding father of the NORSLIS (Nordic Research School in Information Studies), permanent participant and frequent organiser of the Nordic Workshops on Bibliometrics and Research Policy, and amongst others, the two major conferences of the field (ISSI and STI) are also hardly imaginable without a certain corduroy-jacketed Swede with reading glasses carefully placed on his forehead... Could you mention a few of your most memorable conference stories?**
During the first STI conferences (Leiden 1988 and Bielefeld 1990) there were some heated debates on "The decline of British science". The trend seemed to vary significantly depending on a number of methodological decisions, such as constant vs. dynamic journal sets, whole vs. fractional counting etc. This was an early illustration to the need for transparency and controls in developing bibliometric indicators. I remember Kees Le-Pair's urge that we should stop quarrelling in plenum, but rather join in the lab to study the

Wolfgang Glänzel and Olle Persson on the Price Medal ceremony. © Photo copyright: Ann-Britt Persson

Where all the lawn-mowing, wining & dining takes place: summer cottage of the Perssons in Flarken, Sweden. © Photo copyright: Ann-Britt Persson

matters together in a systematic fashion. The debates are still on in our field, but I think we are more open and willing now to critically examine and improve on or main indicators.

I also recall the ISSI 2005 meeting in Stockholm with the evening boat tour in the archipelago. It was bewildering!

A memorable boat tour in the archipelago around Stockholm after the ISSI 2005 Conference organized by Olle Persson and his team. © Photo copyright: Balázs Schlemmer

■ **Rumour has it that one of the NORSLIS summer schools ended with a nice excursion to your very own summer house, where enthusiast PhD students could, amongst others, learn the art of an ancient Swedish tradition, too, in which a rubber-boot plays an inevitable (and flying) role. Apart from this funny game, and the obvious lawn-mowing obligations of a summer cottage owner, what leisure-time activity do you like to do?**

Besides bibliometrics and lawn-mowing I have few talents. I like doing carpentry on my old house since it is already so oblique that it could be combined with two of my other talents: wining and dining.

TEXT MINING AND VISUALIZATION USING VOSVIEWER

NEES JAN VAN ECK
Centre for Science And Technology Studies
Leiden University, the Netherlands
ecknjpvan [at] cwts [dot] leidenuniv [dot] nl

LUDO WALTMAN
Centre for Science and Technology Studies
Leiden University, the Netherlands
waltmanlr [at] cwts [dot] leidenuniv [dot] nl

Abstract: VOSviewer is a computer program for creating, visualizing, and exploring bibliometric maps of science. In this report, the new text mining functionality of VOSviewer is presented. A number of examples are given of applications in which VOSviewer is used for analyzing large amounts of text data.

1. INTRODUCTION

VOSviewer is a computer program that we have developed for creating, visualizing, and exploring bibliometric maps of science (Van Eck & Waltman, 2010). The program is freely available at www.vosviewer.com. VOSviewer can be used for analyzing all kinds of bibliometric network data, for instance citation relations between publications or journals, collaboration relations between researchers, and co-occurrence relations between scientific terms. In this report, we present the new text mining functionality of VOSviewer. We show how this functionality can be used for analyzing large amounts of text data.

2. TEXT MINING FUNCTIONALITY

In September 2011, version 1.4.0 of VOSviewer was released. This new version of VOSviewer includes extensive text mining functionality. The text mining functionality of VOSviewer provides support for creating term maps based on a corpus of documents. A term map is a two-dimensional map in which terms are located in such a way that the distance between two terms can be interpreted as an indication of the relatedness of the terms. In general, the smaller the distance between two terms, the stronger the terms are related to each other. The relatedness of terms is determined based on co-occurrences in documents. These documents can be for instance scientific publications (either titles and abstracts or full texts), patents, or newspaper articles. VOSviewer can only handle English language documents.

To create a term map based on a corpus of documents, VOSviewer distinguishes the following steps:

1. Identification of noun phrases. The approach that we take is similar to what is reported in an earlier paper (Van Eck, Waltman, Noyons, & Buter, 2010). We first perform part-of-speech tagging (i.e., identification of verbs, nouns, adjectives, etc.).

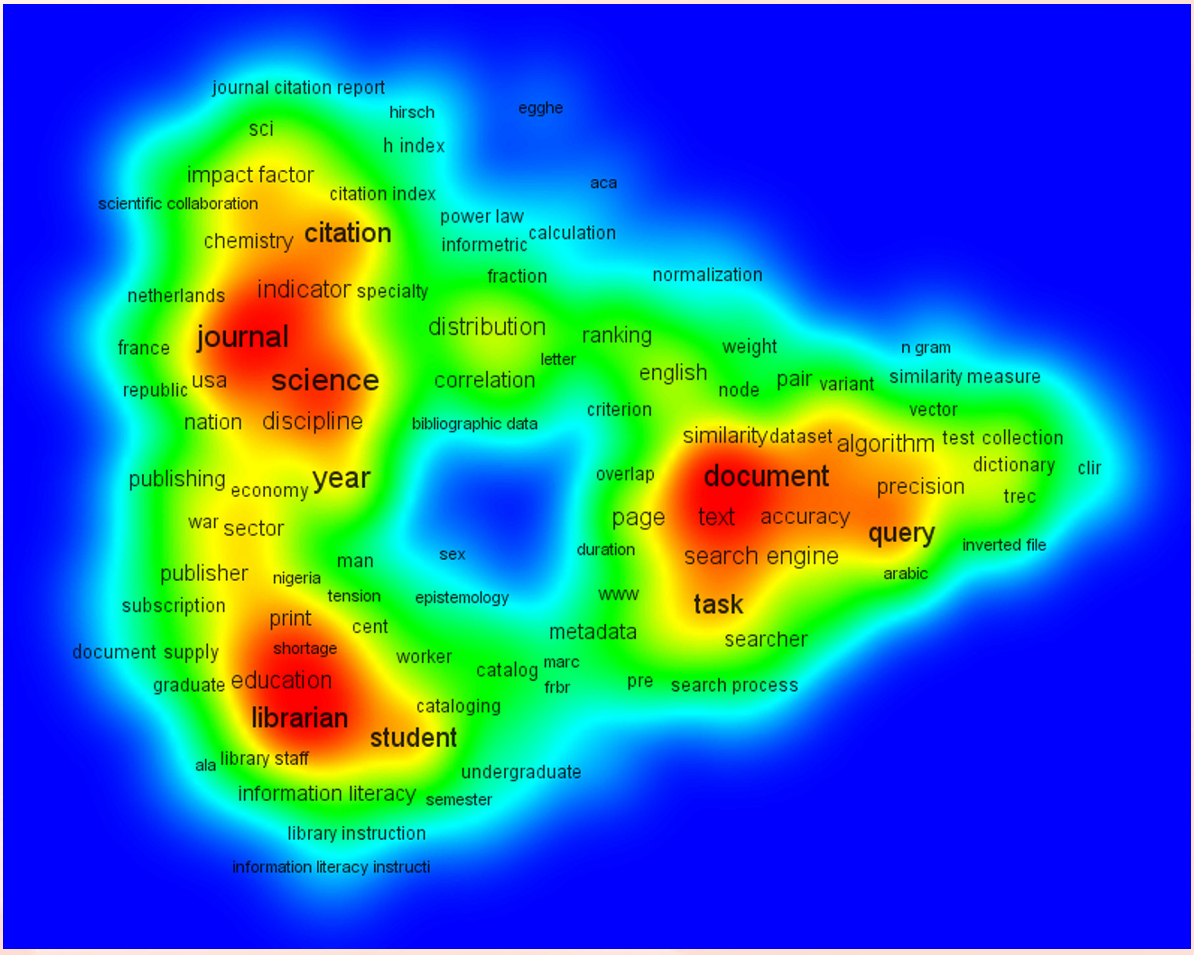


Figure 1. Term map of the field of library and information science. Colors indicate the density of terms.

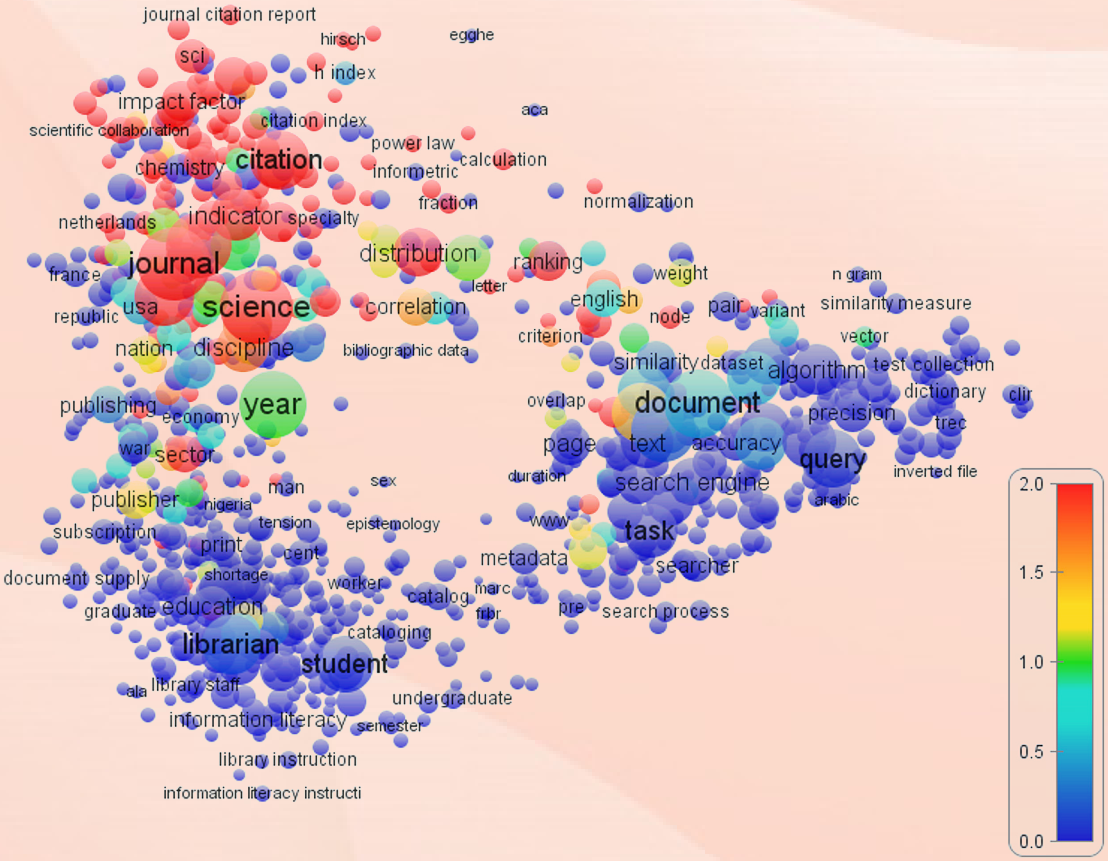


Figure 2. Term map of the field of library and information science. Colors indicate the research activities of Leiden University.

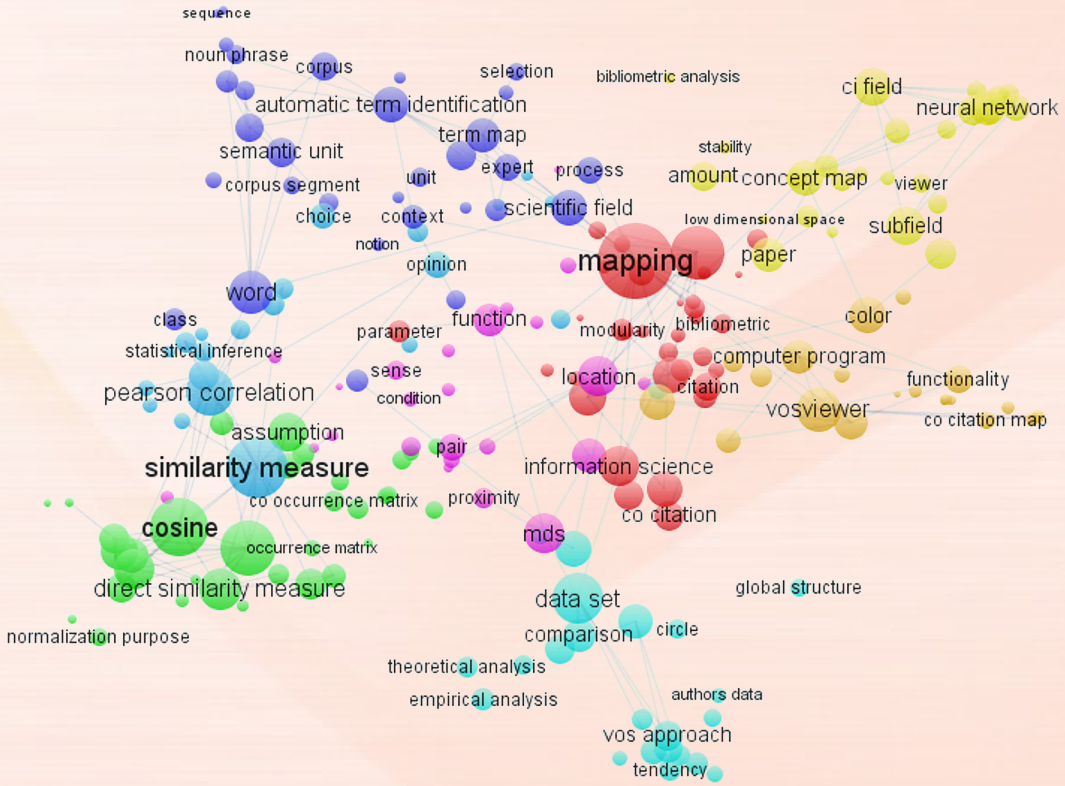


Figure 3. Term map of the full text of Van Eck (2011). Colors indicate clusters of related terms.

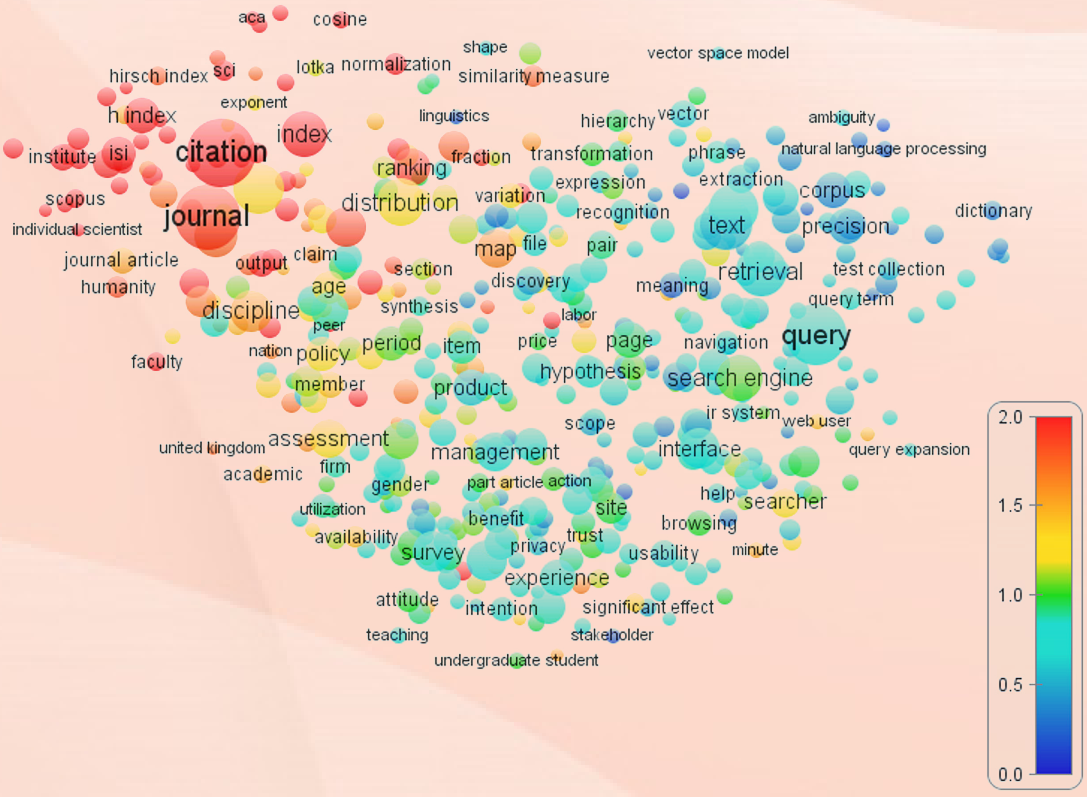


Figure 4. Term map of the Journal of the American Society for Information Science and Technology. The color of a term indicates the average citation impact of the publications in which the term occurs.

The Apache OpenNLP toolkit (http://incubator.apache.org/opennlp/) is used for this purpose. We then use a linguistic filter to identify noun phrases. Our filter selects all word sequences that consist exclusively of nouns and adjectives and that end with a noun (e.g., *paper*, *visualization*, *interesting result*, and *text mining*, but not *degrees of freedom* and *highly cited publication*). Finally, we convert plural noun phrases into singular ones.

2. Selection of the most relevant noun phrases. The selected noun phrases are referred to as terms. We have developed a new technique for selecting the most relevant noun phrases. The essence of this technique is as follows. For each noun phrase, the distribution of (second-order) co-occurrences over all noun phrases is determined. This distribution is compared with the overall distribution of co-occurrences over noun phrases. The larger the difference between the two distributions (measured using the Kullback-Leibler distance), the higher the relevance of a noun phrase. Intuitively, the idea is that noun phrases with a low relevance (or noun phrases with a general meaning), such as *paper*, *interesting result*, and *new method*, have a more or less equal distribution of their (second-order) co-occurrences. On the other hand, noun phrases with a high relevance (or noun phrases with a specific meaning), such as *visualization*, *text mining*, and *natural language processing*, have a distribution of their (second-order) co-occurrences that is significantly biased towards certain other noun phrases. Hence, it is assumed that in a co-occurrence network noun phrases with a high relevance are grouped together into clusters. Each cluster may be seen as a topic.
3. Mapping and clustering of the terms. We use our unified framework for mapping and clustering in this step (Van Eck, Waltman, Dekker, & Van den Berg, 2010; Waltman, Van Eck, & Noyons, 2010).
4. Visualization of the mapping and clustering results. VOSviewer offers various types of visualizations. The program has zoom, scroll, and search functionality to support a detailed examination of a term map.

3. APPLICATIONS

To illustrate the text mining functionality of VOSviewer, we present three examples of applications of this functionality. The term maps that we discuss can be explored in more detail online at www.vosviewer.com/maps/term_maps/.

In the first example, a term map was created based on a corpus of scientific publications in the field of library and information science (LIS). The corpus was extracted from the Web of Science database and consists of the titles and abstracts of about 10,000 publications that appeared in the period 1999–2008 (for more details, see Waltman et al., 2010). Out of the 2101 noun phrases that occur in at least 15 publications in the corpus, the term map contains the 1000 noun phrases that are considered most relevant.

The term map is shown in Figure 1. Colors indicate the density of terms, ranging from blue (lowest density) to red (highest density). As can be seen in Figure 1, examples of prominent terms in LIS research include *journal*, *science*, and *citation*, (upper left), *librarian* and *student* (lower left), and *document*, *task*, and *query* (middle right). These are all single-word terms. Among the slightly less prominent terms, we also observe various multi-word ones, such as *impact factor* (upper left), *information literacy* (lower left), and *search engine* and *test collection* (middle right). The term map also reveals a clear structure of the field. There are three well-separated subfields, which may be referred to as bibliometrics/scientometrics (upper left), library science (lower left), and information science/information retrieval (middle right). The subfields are roughly of equal size. The connection between the bibliometrics subfield and the library science subfield appears to be slightly stronger than the connection of either of these subfields with the information science subfield.

The same term map is also shown in Figure 2. This time colors indicate the research

activities of Leiden University. Colors range from blue to red. A blue term is a term that occurs in no or almost no publications of Leiden University in the field of LIS. A red term is a term that occurs relatively frequently in publications of Leiden University. As expected, the research activities of Leiden University turn out to be strongly focused on the bibliometrics subfield.

A second example of an application of the text mining functionality of VOSviewer is shown in Figure 3. The term map shown in this figure was created based on the full text of a PhD thesis on bibliometric mapping of science (Van Eck, 2011). Each paragraph of the full text was treated as a separate document. The 218 most relevant noun phrases were included in the term map. The colors of the 218 terms indicate clusters of related terms identified by VOSviewer. The clusters turn out to correspond reasonably well with the different chapters of the thesis. For instance, the blue cluster (upper left) represents a chapter on automatic term identification and the orange cluster (middle right) represents a chapter on VOSviewer.

Finally, we consider an application in which the text mining functionality of VOSviewer is used to get some insight into the citation impact of the different topics covered by a journal. The journal that we consider is the *Journal of the American Society for Information Science and Technology* (*JASIST*). The analysis uses data from the Scopus database. Using the text mining functionality of VOSviewer, a term map was created based on the titles and abstracts of all publications that appeared in *JASIST* in the period 2005–2009. The term map contains 468 terms and is shown in Figure 4. The color of a term indicates the average citation impact of the publications in which the term occurs. Colors range from blue (lowest citation impact) to red (highest citation impact). Interestingly, there turn out to be large differences in citation impact among the various topics covered by *JASIST*. The term map indicates a strong separation between bibliometric/scientometric topics on the one hand and information science/information retrieval topics on the other hand. On average, publications on bibliometric/scientometric topics turn out to receive many more citations than publications on information science/information retrieval topics.

4. CONCLUSION

In this report, the new text mining functionality of VOSviewer has been presented. A number of examples have been given of applications in which VOSviewer is used for analyzing large amounts of text data. The examples have focused on scientific texts, but of course the text mining functionality of VOSviewer can also be applied to all kinds of non-scientific texts (e.g., newspaper articles).

We hope that the text mining functionality of VOSviewer will be useful to the bibliometric and scientometric community. We very much welcome feedback from users of our software.

REFERENCES

Van Eck, N.J. (2011). *Methodological advances in bibliometric mapping of science*. PhD thesis, Erasmus University Rotterdam.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Van Eck, N.J., Waltman, L., Dekker, R., & Van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, *61*(12), 2405–2416.

Van Eck, N.J., Waltman, L., Noyons, E.C.M., & Buter, R.K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, *82*(3), 581–596.

Waltman, L., Van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635.

BENFORD'S LAW IS A SIMPLE CONSEQUENCE OF ZIPF'S LAW

LEO EGGHE
Universiteit Hasselt (Uhasselt), Campus Diepenbeek,
Agoralaan, B-3590 Diepenbeek, Belgium*
Universiteit Antwerpen (UA), Stadscampus,
Venusstraat 35, B-2000 Antwerpen, Belgium
leo [dot] egghe [at] uhasselt [dot] be

Abstract: We show that Benford's law (describing the logarithmic distribution of the numbers 1,2,...,9 as first digits of data in decimal form) can be deduced from the classical law of Zipf. This explains Bensford's law as a scientometric or informetric law.
Key words and phrases: Benford's law, Zipf's law
Acknowledgement: The author is grateful to Ronald Rousseau for bibliographical advise on this note.

Benford (1938) rediscovered an earlier finding of Newcomb (1881). Since then it is called Benford's law. This law states that the distribution of the numbers 1,2,...,9 as first digits of data in decimal form is as in (1): the probability to have d = 1,2,...,9 as first digit is

$$P(d) = \log_{10}\left(1+\frac{1}{d}\right) \qquad (1)$$

It is easily seen that P is indeed a distribution:

$$\sum_{d=1}^{9} P(d) = 1 \qquad (2)$$

So the digits 1,2,...,9 are not uniformly distributed as first digits of data in decimal form: (1) shows that smaller digits are favored since (1) is a decreasing function of d.

There are only a few references in the scientometric-informetric literature that mention Benford's law: Brookes and Griffiths (1978) and Brookes (1984) use the name "anomalous law of numbers" and the recent paper Campanario and Coslado (2011). A formal link with existing informetric laws (Lotka, Zipf, ...) has not been given so far. This will be done here. So far, explanations of Benford's law are mathematical - probabilistic - combinatorial and hence occur in the mathematics field - see Cohen (1976), Hill (1995) and Gauvrit and Delahaye (2008) and references therein.

* Permanent address

We will derive Benford's law from the simple law of Zipf

$$g(r) = \frac{B}{r} \qquad (3)$$

where $g(r)$ is the number of item densities in the source on rank density $r \in [1, T]$ (T = total number of sources). Here we will interpret the interval $r \in [d, d+1[$ for $d = 1,2,...,9$ in (3) as the range where the digit d occurs. Hence $T = 10$ here. We first normalise (3) so that it becomes a distribution:

$$\int_1^{10} \frac{B}{r} dr = 1$$

hence

$$B = \frac{1}{\ln(10)}$$

and so (3) reads

$$g(r) = \frac{1}{\ln(10) r} \qquad (4)$$

Then the probability for the digit d to occur ($d = 1,2,...,9$) – hence Benford's law – is given by

$$P(d) = \int_d^{d+1} g(r)\, dr$$

hence

$$P(d) = \frac{1}{\ln(10)} \ln\left(\frac{d+1}{d}\right)$$

or

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \qquad (5)$$

or exactly Benford's law.

A generalization of Benford's law is obtained by replacing (3) by the more general law of Zipf:

$$g(r) = \frac{B}{r^{\beta}} \qquad (6)$$

This generalization (and its practical use) will be studied in a forthcoming paper: Egghe and Guns (2011).

REMARK

Benford's law was recently in the economic "news" in Rauch, Göttsche, Brähler and Engel (2011) were they suggest that (especially) Greece might have tried to make their economic situation seem better. These authors base their finding on statistical deviations of Benford's law. This result was communicated to me by Ronald Rousseau to whom my sincerest thanks.

REFERENCES

F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society 78, 551-572.

B.C. Brookes (1984). Ranking techniques and the empirical log law. Information Processing and Management 20(1-2), 37-46.

B.C. Brookes and J.M. Griffiths (1978). Frequency-rank distributions. Journal of the American Society for Information Science 29, 5-13.

J.M. Campanario and M.A. Coslado (2011). Benford's law and citations, articles and impact factors of scientific journals. Scientometrics 88, 421-432.

D.I.A. Cohen (1976). An explanation of the first digit phenomenon. Journal of Combinatorial Theory (A) 20, 367-370.

L. Egghe and R. Guns (2011). Theory and practise of the generalized law of Benford. Preprint 2011.

N. Gauvrit and J.-P. Delahaye (2008). Pourquoi la loi de Benford n'est pas mystérieuse. Mathematics and Social Sciences 182(2), 7-15.

T.P. Hill (1995). The significant-digit phenomenon. American Mathematical Monthly 102(4), 322-327.

S. Newcomb (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4, 39-40.

B. Rauch, M. Göttsche, G. Brähler and S. Engel (2011). Fact and fiction in EU-Governmental economic data. German Economic Review 12(3), 243-255.

PRELIMINARY FINDINGS ON WHETHER IT IS GOOD VALUE FOR MONEY TO FUND LARGER RESEARCH GROUPS

JONATHAN M. LEVITT
Department of Information Science, Loughborough University, Leicestershire
Statistical Cybermetrics Research Group, University of Wolverhampton
j [dot] levitt [at] lboro [dot] ac [dot] uk and *j [dot] levitt [at] wlv [dot] ac [dot] uk*

Abstract: This poster indicates a way in which citation information can be used to address questions of interest to research policy. Over the last two decades funders and universities have tended to favor research by larger groups of authors. This emphasis has been supported by: (a) the intuitive impression that larger groups are more suited to knowledge- exchange and (b) quantitative evidence that research by larger groups is, on average, more highly cited than research by smaller groups. But a higher average citation does not imply that larger groups are more productive; for example, if larger research groups published a higher proportion of more highly-cited articles, their average citation would be higher irrespective of whether smaller groups published more articles of high citation than did larger research groups. The key finding of this pilot investigation is that for all levels of citation smaller research groups were more productive than larger research groups.
Keywords: information use, co-authorship, citation, research policy

INTRODUCTION AND RELATED RESEARCH

In these current financially stringent times, it is not only important, but also timely, to obtain value for money. This study examines whether it is better value for money to fund research by larger groups of authors than to fund research by small groups of authors.

A recent major goal of research policy is to encourage collaboration. One reason for encouraging collaboration is that the findings of numerous citation studies show that collaborative research tends to be more highly cited than non-collaborative research (e.g., Vogel, 1997; Glänzel, 2000; Glänzel and Schubert, 2001; Leta and Chaimovich, 2002; Goldfinch, Dale, and DeRouen, 2003; Frederiksen, 2004; Uthman, 2008; Levitt and Thelwall, 2009). These studies did not normalize for the number of authors; irrespective of the number of authors, every author of an article was given citation credit for the article.

However, Levitt, Thelwall and Levitt (2011) found that when citations are normalized for the number of authors, the quantity

of articles and citations is inversely related to the level of authorship, in that the number of articles and citations decreases with increased level of authorship (findings summarized in Table 5, the Appendix).

A possible explanation for these contrasting findings is that, although the percentage of less highly-cited articles decreases with increased level of authorship, the number of citations per author also decreases with increased level of authorship. If this is the case, then the implication is that, for UK physics, more highly collaborative researchers are less productive, if productivity is measured by their total citation of the articles published in a year.

This current research further investigates the data used by Levitt, Thelwall and Levitt (2011) to compare the citation profiles of diverse levels of authorship. Specifically, it evaluates for each authorship-level the number of articles in six strata of percentiles of citation that may be expected from 1,000 'average researchers' in the authorship-level. The rationale for this study is that it provides a way of comparing the anticipated research yield from funding researchers of diverse citation levels. Any indication that the yield is not substantially higher for research by larger groups of authors would be of particular interest, as it would raise doubts on the justification for the current emphasis of research funders to encourage collaborative research.

DATA AND METHODS

This research investigates the 3,266 articles published in 2004 in the Science Citation Index (SCI), with at least one author with a UK address, in one or more of the following SCI subject categories: Astronomy & Astrophysics; Physics, Multidisciplinary; Physics, Particles & Fields. The categories were chosen, as they contain more than 480 articles published in 2004 and contain a particularly large number of articles that are each by more than 100 researchers. The minimum size of 480 was set a 480t, in order to ensure an average of at least 6 articles for each of the 80 pairs of author-level and citation strata; Physics, Nuclear was excluded as only 216 articles published in 2004 are in that category. The number of articles published in 2004 in the other subject is presented in Table 5, the Appendix.

AVERAGE LEVEL OF AUTHORSHIP	LESS THAN 60%	TOP 40 TO 60%	TOP 20 TO 40%	TOP 10 TO 20%	TOP 10%	ALL
1 ≤ and < 2	946.4	171.0	125.4	45.6	34.2	1322.7
2 ≤ and < 3	363.7	155.9	154.4	56.3	49.1	779.3
3 ≤ and < 4	219.7	119.5	123.6	73.2	51.0	587.0
4 ≤ and < 5	161.1	93.0	105.0	34.7	39.4	433.3
5 ≤ and < 6	137.0	86.0	76.5	35.4	27.9	362.8
6 ≤ and < 7	122.7	53.7	73.3	26.1	29.4	305.2
7 ≤ and < 8	102.0	63.0	45.6	25.7	18.7	255.1
8 ≤ and < 9	85.3	39.6	48.5	26.2	22.3	221.9
9 ≤ and < 10	75.6	38.8	42.3	19.2	21.2	197.1
10 ≤ and < 11	55.9	45.6	47.4	27.4	20.7	197.0
11 ≤ and < 16	34.1	32.6	37.2	18.7	29.7	152.3
16 ≤ and < 26	18.3	21.4	22.1	17.0	27.1	105.9
26 ≤ and < 51	7.6	17.7	18.2	12.0	19.0	74.5
51 ≤ and < 101	0.2	3.3	4.5	2.8	12.2	23.1
≥ 751	0	0	1.1	0	0	1.1
101 ≤ and < 751	5.0	3.1	4.1	2.6	2.1	16.9
All	2334.7	944.2	929.2	423.0	403.9	5034.9

Table 1: Article output in diverse citation strata from a hypothetical 1000 'average researchers' for 'Astronomy & Astrophysics'.

AVERAGE LEVEL OF AUTHORSHIP	LESS THAN 60%	TOP 40 TO 60%	TOP 20 TO 40%	TOP 10 TO 20%	TOP 10%	ALL
1 ≤ and < 2	838.9	232.0	160.6	53.5	35.7	1320.8
2 ≤ and < 3	302.0	120.1	157.0	46.4	30.9	656.4
3 ≤ and < 4	185.1	82.9	86.3	37.3	32.5	424.1
4 ≤ and < 5	81.0	65.1	78.1	48.4	36.9	309.5
5 ≤ and < 6	58.7	49.2	45.2	21.4	52.4	227.0
8 ≤ and < 9	32.6	25.0	38.4	14.4	24.0	134.4
6 ≤ and < 7	30.0	40.9	57.6	22.5	40.9	191.9
7 ≤ and < 8	28.1	30.5	53.9	28.9	24.1	165.6
10 ≤ and < 11	25.6	11.5	33.3	6.4	30.7	107.6
9 ≤ and < 10	21.0	45.8	7.6	15.3	47.7	137.5
11 ≤ and < 16	14.1	25.8	19.8	33.2	10.6	103.6
16 ≤ and < 26	5.2	19.2	29.9	6.4	3.0	63.7
26 ≤ and < 51	4.8	2.6	18.9	9.7	3.5	39.5
101 ≤ and < 751	4.6	4.8	3.9	2.1	4.1	19.6
51 ≤ and < 101	.3	.1	8.3	5.1	.4	14.3
≥ 751	0	0	0	0	0	0
All	1632.2	755.6	798.8	351.3	377.5	3915.4

Table 2: Article output in diverse citation strata from a hypothetical 1000 'average researchers' for 'Physics, Multidisciplinary'.

AVERAGE LEVEL OF AUTHORSHIP	LESS THAN 60%	TOP 40 TO 60%	TOP 20 TO 40%	TOP 10 TO 20%	TOP 10%	ALL
1 ≤ and < 2	684.9	185.6	170.3	82.3	86.1	805.4
2 ≤ and < 3	281.2	107.2	125.1	74.8	62.6	530.6
3 ≤ and < 4	160.9	79.6	75.1	30.9	39.8	321.2
4 ≤ and < 5	95.8	37.8	46.8	27.0	28.8	226.7
5 ≤ and < 6	86.4	13.9	27.8	0	41.7	166.7
6 ≤ and < 7	83.3	18.5	104.9	61.7	0	222.2
7 ≤ and < 8	37.0	23.0	30.7	30.7	38.3	214.7
8 ≤ and < 9	92.0	62.7	23.5	62.7	0	164.7
9 ≤ and < 10	15.7	32.1	39.3	0	0	100.0
10 ≤ and < 11	28.6	17.5	8.1	6.7	2.0	88.8
11 ≤ and < 16	54.5	16.7	10.0	6.6	7.8	55.0
16 ≤ and < 26	13.9	2.2	6.2	4.9	3.1	29.9
26 ≤ and < 51	13.5	1.1	4.5	1.3	3.4	16.8
51 ≤ and < 101	6.5	5.0	5.0	3.8	3.8	23.2
101 ≤ and < 751	4.8	4.3	4.6	3.3	3.2	20.2
≥ 751	0	0	1.1	0	0	1.1
All	1659.0	886.5	988.2	457.6	342.5	4333.8

Table 3: Article output in diverse citation strata from a hypothetical 1000 'average researchers' for 'Physics, Particles & Fields'.

It first calculates, for diverse levels of authorship, the number of articles published in six citation strata. It then uses the number of articles in each citation stratum that may be expected from 1,000 'average researchers' in that level. The strata were chosen in such as a way as to provide fine-grained results that were not rendered unreliable because of small sample sizes.

The findings are likely to be skewed towards higher levels of authorship. Firstly, it would be impractical to exclude self-citation

and, in general, research by larger groups of authors is more prone to self-citation. Secondly, and possibly more importantly, research by larger groups seems generally more likely to be more highly funded. Thirdly, capable, well-connected researchers seem likely to be more prevalent in larger research groups. The rationale for this latter assertion is that larger research groups are currently encouraged both by funding bodies and universities.

FINDINGS

Tables 1, 2 and 3 present the preliminary findings of this research. In these tables the rows provide information on diverse levels of authorship, with 1 being assigned to the author of a single author papers, 2 assigned to both authors of a papers with 2 authors, etc. In both tables the columns denote citation level, with 'Less than 60%' denoting articles in the lowest citation level (lowest 40%) and 'Top 10%' denoting articles in the highest citation level.

DISCUSSION AND CONCLUSION

The key findings of Tables 1, 2 and 3 are summarised in Table 4.

Table 4 indicates that even amongst the three citation categories with higher than mean citation, bands with low levels of authorship, in general, have the highest productivity by the hypothetical 1,000 authors; for 8 out of 9 instances the highest productivity is by the bands with average authorship of less than 3.2. This is despite the likelihood that the findings are likely to be skewed towards higher levels of authorship, because of self-citation and sample selection (described in the Methods).

One limitation is that, particularly for the authorship-level of 1 to 2, the findings are based on a small sample. A second limitation is that the findings are confined to UK articles. A third limitation is that this study investigates a subset of the SCI physics categories. We will address these limitations by expanding the study to all articles published in 2004 that are in SCI physics categories. A fourth limitation is that the findings are based on a single subject. We will address this limitation by investigating other subjects.

Despite these limitations the findings are interesting as they cast doubt on the current practice of funders and academia encouraging co-authorship.

Any doubt on the merits of encouraging co-authored research is highly relevant to research policy, as: (1) Especially in the current financial stringency, it seems inappropriate to devote resources to research by larger groups if it is not clear-cut that it is likely to produce a higher yield for a given number of funded researchers; and (2) Articles by large research groups may hamper precise evaluation of research, irrespective of the method of evaluation. For instance, for a multi-authored article an outsider would have problems assessing precisely the contribution by different authors to the article. This could result in a less precise assessment of the research contribution of individuals, research groups, departments, countries and funded research. This final point applies not only to research assessment using bibliometrics, but also to research assessment using peer review.

SUBJECT	LESS THAN 60%	TOP 40 TO 60%	TOP 20 TO 40%	TOP 10 TO 20%	TOP 10%
Astronomy & Astrophysics	1 ≤ and < 2 (av. 1.08)	1 ≤ and < 2	2 ≤ and < 3 (av. 2.11)	3 ≤ and < 4 (av. 3.11)	3 ≤ and < 4
Physics, Multidisciplinary	1 ≤ and < 2 (av. 1.08)	1 ≤ and < 2	1 ≤ and < 2	1 ≤ and < 2	5 ≤ and < 6 (av. 5.02)
Physics, Particles & Fields	1 ≤ and < 2 (av. 1.11)	1 ≤ and < 2	1 ≤ and < 2	1 ≤ and < 2	1 ≤ and < 2

Table 4: For diverse citation percentiles, the author level with the highest level of productivity (average authorship in band in brackets).

ACKNOWLEDGEMENT

This research is funded by the Economic and Social Research council, Grant Reference: RES-000-22-4415. I would like to thank Gertrude Levitt and Mike Thelwall for their helpful feedback.

REFERENCES

Frederiksen L.F. (2004). Disciplinary determinants of bibliometric impact in Danish industrial research: Collaboration and visibility. *Scientometrics*, 61(2), 253-270.

Glänzel, W. & Schubert A. (2001). Double effort = Double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50(2), 199-214.

Glänzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, 48(2), 121-150.

Goldfinch S., Dale T. & DeRouen K. (2003). Science from the periphery: Collaboration, networks and 'Periphery Effects' in the citation of New Zealand Crown Research Institutes articles, 1995-2000. *Scientometrics*, 57(3), 321-337.

Leta J. & Chaimovich H. (2002). Recognition and international collaboration: the Brazilian case. *Scientometrics*, 53(3) 325-335.

Levitt J.M. & Thewall M. (2009). Citation levels and collaboration within Library and Information Science. *Journal of the American Society for Information Science and Technology*, 60(3), 434-442.

Levitt J.M., Thewall M. & Levitt M. (2011). *Proceedings of the 13th Conference of the International Society for Scientometrics and Informetrics*, volume 1, 398-408.

Vogel, E.E. (1997). Impact factor and international collaboration in Chilean physics: 1987-1994. *Scientometrics*, 38(2), 253-263.

APPENDIX

Table 6 is relegated to the Appendix, as it reproduces the findings of a conference presentation (Levitt, Thelwall and Levitt, 2011). It is included in this proposal, as this current proposal builds on this table and it demonstrates that the findings of this current proposal extend considerably on the previously published study.

SUBJECT	60TH	40TH	20TH	10TH	ARTICLES
Astronomy & Astrophysics	12	20	38	57	1874
Physics, Multidisciplinary	8	16	36	58	1054
Physics, Particles & Fields	9	16	32	51	733
Physics, Nuclear	9	14	24	34	216
Not only Physics, Nuclear	10	18	36	56	3266
All subjects	10	18	36	55	3364

Table 5: Citation percentiles and number of articles.

AVERAGE LEVEL OF AUTHORSHIP	SAMPLE SIZE	ARTICLES PER RESEARCHER	CITATIONS PER RESEARCHER
1 ≤ and < 2	181	1.40	8.76
2 ≤ and < 3	765	.78	8.15
3 ≤ and < 4	1,057	.55	6.51
4 ≤ and < 5	936	.42	6.01
5 ≤ and < 6	766	.35	5.88
6 ≤ and < 7	566	.28	4.34
7 ≤ and < 8	534	.26	3.61
8 ≤ and < 9	362	.22	2.98
9 ≤ and < 10	298	.21	3.30
10 ≤ and < 11	294	.20	2.85
11 ≤ and < 12	224	.23	2.50
12 ≤ and < 13	172	.19	4.31
13 ≤ and < 14	206	.16	2.45
14 ≤ and < 15	174	.14	2.93
15 ≤ and < 16	155	.15	1.89
16 ≤ and < 17	100	.17	2.17
17 ≤ and < 18	124	.15	1.85
18 ≤ and < 19	119	.15	1.92
19 ≤ and < 20	131	.13	1.68
1 ≤ and < 30	8,349	.36	4.49
30 ≤ and < 60	1,033	.08	1.19
60 ≤ and < 90	486	.05	1.85
90 ≤ and < 120	422	.04	.57
120 ≤ and < 150	544	.03	.61
150 ≤ and < 180	93	.09	2.17
180 ≤ and < 210	551	.03	.44
210 ≤ and < 240	505	.06	.27
240 ≤ and < 270	71	.11	.32
270 ≤ and < 300	301	.03	.09
300 ≤ and < 330	446	.05	.16
330 ≤ and < 360	1,141	.04	.30
360 ≤ and < 390	757	.02	.20
390 ≤ and < 420	28	.06	.19
420 ≤ and < 450	26	.06	.09
450 ≤ and < 480	55	.06	.19
480 ≤ and < 510	348	.04	.10
510 ≤ and < 540	268	.04	.08
540 ≤ and < 570	188	.04	.06
600 ≤ and < 630	380	.07	.08

Table 5: Average article and citation productivity for diverse levels of authorship for almost all authors of UK physics 2004 articles.*

** The 45 individuals with an average level of authorship of over 630 were excluded because more than 80% (37) authored the same single article.*

BANGLADESH' PUBLICATION BARYCENTRE

DILRUBA MAHBUBA
275, New Ailpara, Pathantuli, Shiddhirgonj Narayangonj-1412, Bangladesh
Antwerp University, IBW, Belgium
dilrubaauw [at] gmail [dot] com

RONALD ROUSSEAU
KHBO (Association K.U.Leuven), Faculty of Engineering Technology, Oostende, Belgium
K.U.Leuven, Dept. Mathematics, Leuven (Heverlee), Belgium
ronald [dot] rousseau [at] khbo [dot] be

Abstract: Using the barycentre method we represent and visualize the relative publication change among regions in Bangladesh and illustrate the important role played by its capital Dhaka. Data for this study are obtained from Thomson Reuters' Web of Science (March 2011). We found that Bangladesh' publication barycentre moves slowly and is situated to the East of the capital, Dhaka. Inequality of publication is very large among districts, showing a high concentration in a few places, mainly in the capital.

INTRODUCTION

The term barycentre comes from the Greek word *βαρύκεντρον*. The prefix "bar" is a Greek root meaning weight or heavy. Literally, the word barycentre means the "centre of weight." The concept of a barycentre (defined precisely further on) is also known as the centre of gravity, centre of weight, orthocentre, centroid, mass centre, or centre of mass, depending on the practical context. The idea, originating from Euclidean geometry (Lay, 2012, p. 33), has been adapted and applied in many fields, such as physics, astronomy, statistics and demography.

Bartlett (1985) applied elementary concepts of kinematics and Newtonian dynamics of an ensemble of particles to the population of the U.S. in analysing the motion of the centre of population (centre of mass). It is well-known that over the years the centre of the American population has moved to the West. Rousseau (1989a) introduced kinematical statistics of scientific output and the barycentre method in informetrics representing results on an actual geographic map. As an illustration he determined the publication barycentre or the centre of publication (CPub) of the scientific output of the Scandinavian countries (Rousseau, 1989a). Jin & Rousseau (2001) studied China's publication centre and its movement over a period of ten years. They illustrated that, slowly, China's science became less centralized, i.e.

less dominated by institutes and universities located in Beijing. For cases where a geographical map makes no sense Rousseau (1989b) considered an abstract representation in a regular polygon, see also (Egghe & Rousseau, 1990) and (Rousseau, 2008).

Our study provides a visual geographical representation of the centre of Bangladeshi publications and its change over four decades from 1971 to 2010. Data are collected and analysed separately for the four decades: 1971-1980; 1981-1990; 1991-2000 and 2001-2010. By comparing their values we find how the barycentre has moved.

MATHEMATICAL DEFINITION OF THE CENTRE OF PUBLICATION

The centre of scientific publication, or publication barycentre, of a country (or any other geographical unit) is defined as the point at which a flat, weightless but stiff map of the country would balance if weights of identical value were placed on it so that each weight represented the origin of one publication. This centre of publication

$$\vec{C} = \left(C_x, C_y\right)$$

is calculated according to the following formulae:

$$C_x = \frac{\sum_{i=1}^{n} m_i \cdot L_{i,x}}{M} \quad , \quad C_y = \frac{\sum_{i=1}^{n} m_i \cdot L_{i,y}}{M} \qquad (1)$$

where n is the number of elements in the system,

$$\vec{L_i} = \left(C_{i,x}, C_{i,y}\right)$$

is the location of the i^{th} element in the system and m_i is the contribution (here: number of publications) of the i^{th} element (during a fixed period). Further,

$$M = \sum_{i=1}^{n} m_i$$

denotes the total number of publications in the system; m_i / M is the relative contribution of the i^{th} element.

We study Bangladeshi publications to illustrate this approach.

BANGLADESH

Bangladesh, one of the highly populated neighbouring countries of India (Mahbuba & Rousseau, 2010), is situated near latitude 24° north and longitude 90° east. Dacca, respelled Dhaka, which is the capital city of Bangladesh, is situated at latitude 23° 51' north and longitude 90° 24' east. Another major city of Bangladesh, named Chāṭṭagrām, respelled Chittagong is located at 22° 21' north latitude and 91° 50' east longitude. Bangladesh has 64 administrative divisions called districts. The population density of its capital Dhaka is the highest with 8,625 inhabitants per km². Because of its function and its density it is no surprise that a major portion of scientific publications is produced in the capital. Longitude and latitude of the main publication producing districts and their population density per km² are shown in Table 1.

DATA COLLECTION

In March 2011 we performed advanced searches in the WoS using all databases, namely SCI-Expanded, SSCI, A&HCI, CPCI-S, and CPCI-SSH. The query AD = Bangladesh with time span = 1971-2010 led to a total of 14,946 publications. This search was actually performed for each of the four decades 1971-1980, 1981-1990, 1991-2000, 2001-2010 separately. In this way we were able to calculate four barycentres and to study their geographical changes.

Figure 1 shows the institutes that contribute most to the scientific output of Bangladesh, and their location within the country.

The twenty most productive Bangladeshi institutes during this period, as found by our search, are shown in Table 2.

Bangladeshi scientific publications (ScPub) come mainly from Dhaka, Chittagong, Gazipur, Mymensingh, Rajshahi, Sylhet and Khulna. We found 84 different institutions (with at least two publications in the WoS) contributing to the country's

DISTRICTS	LATITUDE (DECIMAL DEGREE)	LONGITUDE (DECIMAL DEGREE)	POPULATION DENSITY (PER KM²)	AREA (KM²)	ANNUAL POPULATION GROWTH (IN %)
Dhāka	23.70°N	90.39°E	8 625	1 439	3.64
Chāṭṭagrām	22.33°N	91.81°E	1 644	4 866	1.71
Gazipur	23.80°N	90.65°E	1 445	1 798	2.25
Maimansingh	24.75°N	90.39°E	1 181	4 399	1.22
Rājshāhī	24.37°N	88.59°E	1 139	2 426	1.7
Silhaṭ	24.90°N	91.87°E	939	3 312	1.6
Khulnā	22.84°N	89.56°E	581	4 746	1.33

Table 1. The seven scientifically most-active districts of Bangladesh
Source: http://www.world-gazetteer.com [downloaded March 2011]

scientific output. Among these, 57 institutions are situated in the district of Dhaka. These Dhaka institutions have published 67.22 percent (N=10,047: 9140 in Dhaka City; 907 in Savar) of the country's output of a total of 14,946 publications (until 2010).

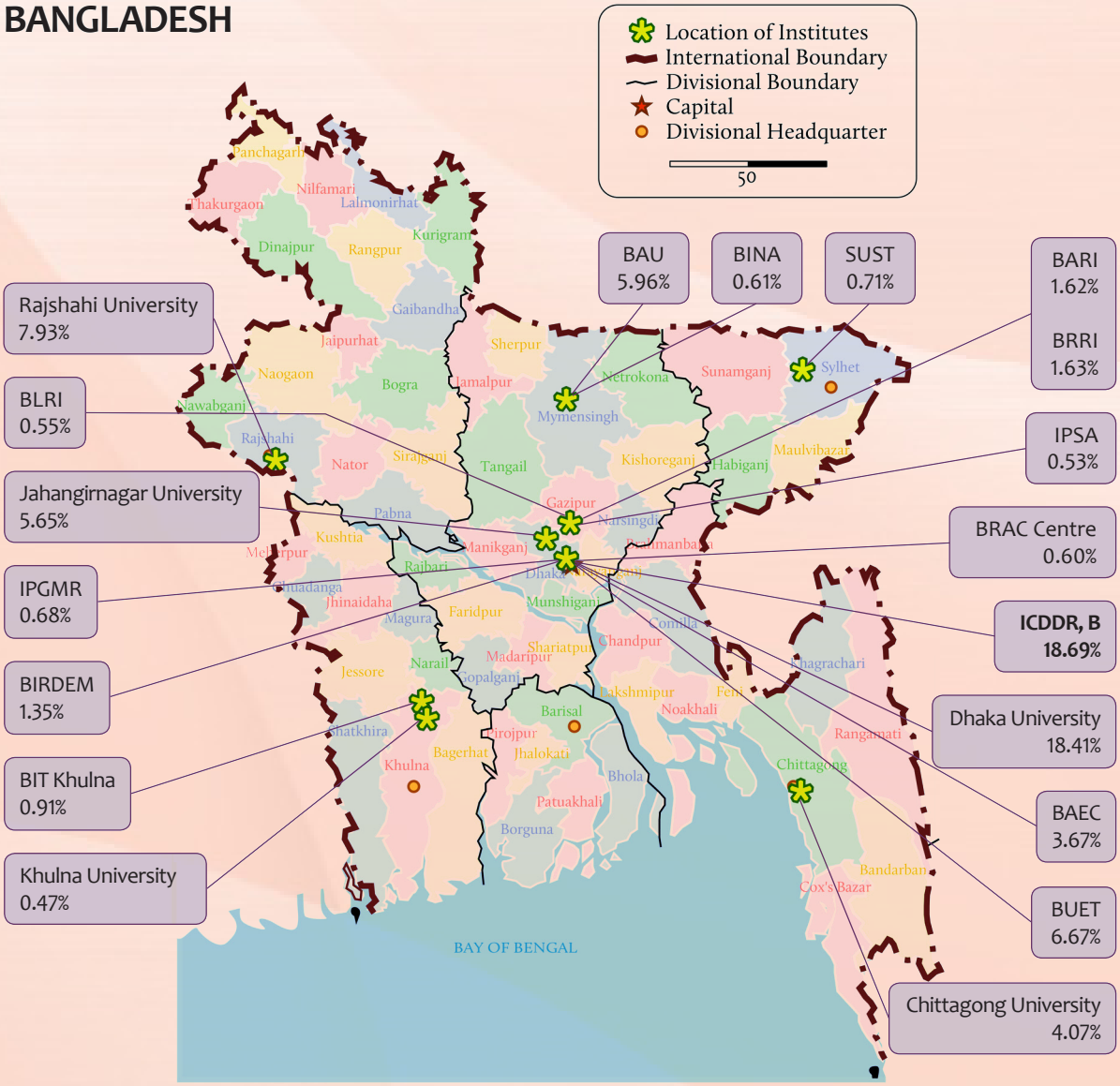


Figure 1. Most contributing institutions (ICDDR,B map).
Source of base map: Wikimedia Commons. Copyright by Armanaziz, licensed under the Creative Commons Attribution 2.5 Generic license (http://creativecommons.org/licenses/by/2.5/deed.en)

INSTITUTES	TOTAL NUMBER OF DOCS	PERCENTAGE
International Centre for Diarrhoeal Disease Research, Bangladesh	2684	17.96
University of Dhaka	2450	16.39
Bangladesh University of Engineering and Technology	1390	9.30
Rajshahi University	1172	7.84
Bangladesh Agricultural University	820	5.49
Jahangirnagar University	809	5.41
Bangladesh Atomic Energy Commission	576	3.85
Chittagong University	559	3.74
Bangladesh Agriculture Research Institute	292	1.95
Shahjalal University of Science and Technology	276	1.85
Bangladesh Rice Research Institute	228	1.53
Khulna University of Engineering and Technology	222	1.49
Bangladesh Institute of Research and Rehabilitation for Diabetes, Endocrine and Metabolic Disorders	215	1.44
Bangladesh Council for Science and Industrial Research	162	1.08
Bangabandu Sheikh Mujib Medical University (formerly: Institute of Postgraduate Medical Research)	154	1.03
Bangladesh Rural Advancement Commission	150	1.00
Dhaka Medical College Hospital	144	0.96
Dhaka Children's Hospital	135	0.90
Bangladesh Institute of Nuclear Agriculture	103	0.69
East West University	94	0.63

Table 2. The twenty most productive Bangladeshi institutes (according to WoS data)

DATA COLLECTION AND FILTERING

To calculate the barycentre of Bangladeshi publications we used the following data:

a) The origin (address) of the contributing organization
b) The geographic coordinates (latitude and longitude) of this organization
c) The number of contributed publications

Data cleaning was necessary as we found many name variations. As an example of such a name variation we mention the University of Dhaka which can be found as:

DACCA UNIV 34
DHAKA UNIV 69
UNIV DACCA 246
UNIV DHAKA 2099

For ICDDR,B (International Centre for Diarrhoeal Diseases Research, Bangalore) we even found 27 name variants. Next we tried to find geographic coordinates (GC) as precisely as possible. First we collected all addresses (geographic location from institutions' websites) from Google Earth for precise locations; "dms" style latitudes and longitudes were converted to decimal latitudes and longitudes.

THE CENTRE OF PUBLICATION (CENTRE OF MASS): ITS CALCULATION

According to the formulae shown above (1) we calculated the publication barycentre of Bangladesh per decade and tried to find if there was a systematic movement over different decades. Results are shown in Table 3.

These barycentres move up and down somewhat to the East of Dhaka. However, no systematic change is visible (see Fig.2).

PERIOD	C_x	C_y
1971-1980	23.7900	90.3254
1981-1990	23.8533	90.3233
1991-2000	23.8024	90.3131
2001-2010	23.8209	90.3203

Table 3. Coordinates of barycentres

DISCUSSION

Different types of organizations including research institutes, universities, colleges, NGOs, government organizations, a total of 576, contributed to the publication output of Bangladesh, as covered by the WoS. Most institutes occurred with different names (in the WoS) so that name disambiguation was necessary. For instance, 81 institute names were refined to 18 unique ones.

The top twenty institutes (this is 3.5% of all institutes) published 84.53 % or 12,635 publications. This points to a situation which is much more concentrated than expected from the proverbial 80/20 rule. Let us have a look at their research output and their geographic positions in Table 4.

DEVELOPMENTS

During the first decade Dhaka University was the most active, but the following three

	INSTITUTES	PUB-LICA-TIONS	% OF TOTAL (14,946)	PLACE	Y_LAT	X_LON
1	International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B)	2684	17.96	Mohakhali, Dhaka	23.776442	90.399789
2	University of Dhaka (DU)	2450	16.39	Ramna, Dhaka	23.734192	90.392828
3	Bangladesh University of Engineering and Technology (BUET)	1390	9.30	Bakshibazar, Dhaka	23.726372	90.392469
4	Rajshahi University (RU)	1172	7.84	Rajshahi	24.372572	88.637289
5	Bangladesh Agricultural University (BAU)	820	5.49	Mymenshingh	24.722917	90.429633
6	Jahangirnagar University (JU)	809	5.41	Savar, Dhaka	23.877039	90.268311
7	Bangladesh Atomic Energy Commission (BAEC)	576	3.85	Kazi Nazrul Islam Avenue, Dhaka	23.763252	90.389113
8	Chittagong University (CU)	559	3.74	Chittagong	22.471078	91.786664
9	Bangladesh Agriculture Research Institute (BARI)	292	1.95	Gazipur	23.994878	90.416203
10	Shahjalal University of Science and Technology (SUST)	276	1.85	Sylhet	24.923956	91.832547
11	Bangladesh Rice Research Institute (BRRI)	228	1.53	Gazipur	23.991647	90.409011
12	Khulna University of Engineering and Technology (KUET)	222	1.49	Fulbarigate, Khulna	22.900583	89.501544
13	Bangladesh Institute of Research and Rehabilitation for Diabetes, Endocrine and Metabolic Disorders (BIRDEM)	215	1.44	Shahbagh, Dhaka	23.738761	90.396611
14	Bangladesh Council for Science and Industrial Research (BCSIR)	162	1.08	Science Laboratory, Dhaka	23.739439	90.384558
15	Bangabandu Sheikh Mujib Medical University (BSMMU)	154	1.03	Shahbag, Dhaka	23.739003	90.395044
16	Bangladesh Rural Advancement Commission (BRAC)	150	1.00	Mohakhali, Dhaka	23.714897	90.406564
17	Dhaka Medical College Hospital (DMCH)	144	0.96	Dhaka	23.776017	90.398264
18	Dhaka Children's Hospital (DCH)	135	0.90	Sher-e-Bangla Nagar, Dhaka	23.77175	90.37719
19	Bangladesh Institute of Nuclear Agriculture (BINA)	103	0.69	Mymenshingh	24.749353	90.399667
20	East West University (EWU)	94	0.63	Mohakhali, Dhaka	23.780644	90.407353

Table 4: Research output of top 20 institutes and their geographic positions

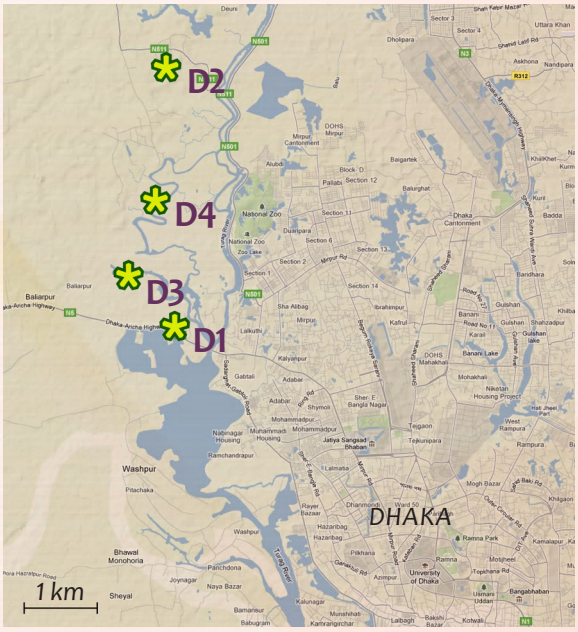


Fig 2. Barycentres over decades shown on a Google map using GPSVisualizer. Map data © 2011 Google. Map created at © GPSVisualizer.com

ones ICDDR,B took over the lead. Jahangirnagar University situated in Savar, in the east of Dhaka occupies a strong position over all decades. As the percentage of relative contributions varied little over different decades the barycentre also changed little (and always stayed somewhat at the east of Dhaka City).

CONCLUSION

We showed how Bangladesh' barycentre of scientific publications evolved during the period 1971-2010 and presented a geographical approach of its position and movement over four decades. The barycentre method was used on publication data. Yet, this approach can also be applied to various other types of data for useful visualization. Further study with public health related data (work in progress) may reveal interesting results which can help in policy making issues in the country.

ACKNOWLEDGEMENT

We thank Mr. Muhammad Zahirul Haq, Data Management Supervisor, Health and Demographic Surveillance Unit, ICDDR,B for his contribution in making Fig. 1 using ArcGIS software.

REFERENCES

Bartlett, A.A. (1985). U.S. population dynamics. *American Journal of Physics* 53, 242-48

Egghe, L. & Rousseau, R. (1990). *Introduction to Informetrics. Quantitative methods in library documentation and information science*. Elsevier: Amsterdam.

Jin, BH. & Rousseau, R. (2001). An introduction to the barycentre method with an application to China's mean centre of publication. *Libri*, 51, 225-233.

Lay, D.C. (2012). *Linear algebra and its applications*. Addison-Wesley: Boston.

Mahbuba, D. & Rousseau, R, (2010). Scientific research in the Indian subcontinent: selected trends and indicators 1973-2007: comparing Bangladesh, Pakistan and Sri Lanka with India, the local giant. *Scientometrics*, 84(2): 403-420.

Rousseau, R. (2008). Triad or tetrad: another representation. *ISSI Newsletter*, 4(1), 5-7.

Rousseau, B. & Rousseau, R.. (2000). LOTKA : a program to fit a power law distribution to observed frequency data. *Cybermetrics* 4: paper 4. http://www.cindoc.csic.es/cybermetrics/articles/v4ilp4.html

Rousseau, R. (1989a). Kinematical statistics of scientific output. Part I: geographical approach. *Revue française de bibliométrie*, 4, 50-64.

Rousseau, R. (1989b). Kinematical statistics of scientific output. Part II: standardized polygonal approach. *Revue française de bibliométrie*, 4, 65-77.

World Gazetteer. http://www.world-gazetteer.com [viewed March 2011]