Demographic Research: Volume 30, Article xx
Descriptive Finding

Quantifying paradigm change in demography

Jakub Bijak[1]

Daniel Courgeau[2]

Eric Silverman[3]

Robert Franck[4]

Abstract

BACKGROUND
Demography is a uniquely empirical research area amongst the social sciences. We posit that the same principle of empiricism should be applied to studies of the population sciences as a discipline, contributing to greater self-awareness amongst its practitioners.

OBJECTIVE
The paper aims to include measurable data in the studies of changes in selected demographic paradigms and perspectives.

METHODS
The presented analysis is descriptive and is based on a series of simple measures obtained from the free online tool *Google Books Ngram Viewer*, which includes frequencies of word groupings (*n*-grams) in different collections of books digitised by Google.

RESULTS
The tentative findings corroborate the shifts in the demographic paradigms identified in the literature – from cross-sectional, through longitudinal, to event-history and multilevel approaches.

[1] Social Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom.
E-Mail: j.bijak@soton.ac.uk
[2] Research Director Emeritus, Institut national d'études démographiques (INED), Paris, France.
[3] University of Southampton, United Kingdom.
[4] Professor Emeritus, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

CONCLUSIONS
The findings identify a promising area of enquiry into the development of demography as a social science discipline. We postulate that more detailed enquiries in this area in the future could lead to establishing History of Population Thought as a new sub-discipline within population sciences.

1. Introduction

The year 2012 marked the 350th anniversary of the publication of John Graunt's *Bills of Mortality* and – arguably – the birth of demography as a formal discipline of scientific enquiry. In accordance with the long-standing empirical tradition of demography as a standalone research area within social sciences (Morgan and Lynch 2001; Courgeau 2012), this paper aims to include measurable data in the studies of the changes in demographic paradigms and theories. After Courgeau and Franck (2007), and following the original suggestions of Granger (1994), we interpret paradigms as studies of different 'scientific objects'. To study their dynamics, we propose using the free online tool, *Google Books Ngram Viewer*.

This paper is entirely devoted to presenting and interpreting selected descriptive findings from the paradigmatic quest mentioned above, and is therefore structured as follows. After this Introduction, we illustrate our argument in Section 2 by using examples related to the demographic nomenclature, studies of different components of demographic dynamics, and to theoretical and paradigmatic change in demography. Section 3 contains a discussion of selected findings, followed by a brief evaluation of some of the potential benefits and limitations of the application of the proposed method. We conclude by proposing an open challenge for the demographic community in Section 4, related to establishing the History of Population Thought as a fully formed sub-discipline of population sciences.

2. Demographic paradigms and *n*-gram analysis: Principles and illustrations

As proposed by Courgeau and Franck (2007: 44), the successive paradigms of demography "describe the various types of relationship between the phenomena observed and the scientific object", whereby the object of scientific interest is the change of human populations. The four paradigms proposed by Courgeau and Franck (2007) – cross-sectional, longitudinal, event-history, and multilevel – are thus related to

the changing and mutually complementary perspectives through which the relationships between population parameters, and between individuals and populations, are being examined. Still, even 350 years after its inception, demography is thought to be a "science in the making", in need of a more solid grounding through axiomatisation (idem). Potential further developments also include theory building – something that is seen as one of the key challenges of contemporary population sciences (see e.g., the discussion in Xie 2000 and Burch 2003). The analysis of changes in existing paradigms and the development of new ones can bring demography closer to achieving these aims.

On the other hand, demography is renowned amongst social science disciplines for being, for the most part, a thoroughly empirical area of enquiry. This is considered to be the main source of the past successes of population studies, alongside the practical applications of research results in the public policy field (for a discussion, see e.g., Xie 2000, and Morgan and Lynch 2001). In addition, demographic works are also on average cited more frequently than those in other social science disciplines (van Dalen and Henkens 2001). Even though there is a gap between different publication venues (idem), and citation rates in population sciences as such do not allow for complacency, this can be seen as a sign of a healthy exchange of ideas. Given these dynamics, demography offers a quite unique testing ground for a quantitative analysis of the changes in its paradigms and theories.

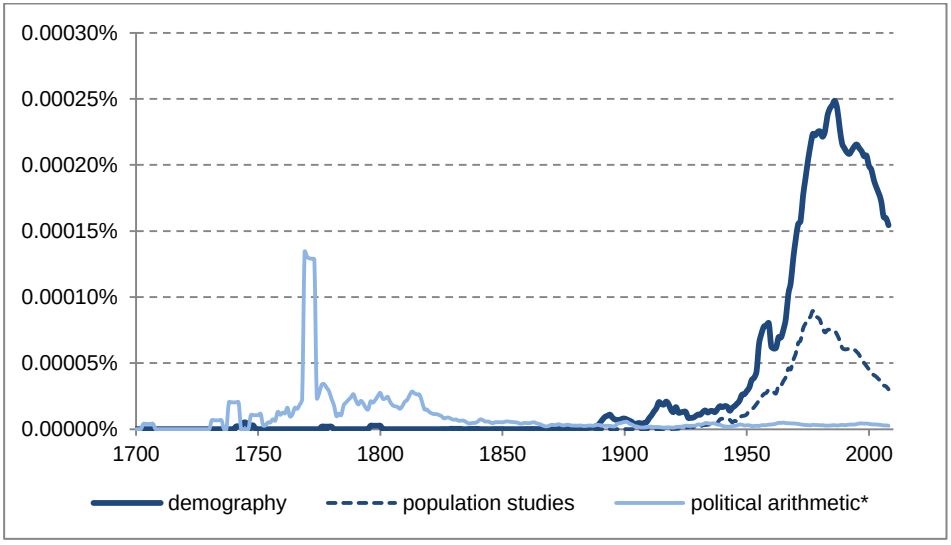
In this paper we follow the previous examples of quantitative content analysis of demographic literature (e.g., Teachman, Paasch, and Carver 1993; Keyfitz 1993; van Dalen and Henkens 2001). The illustrations presented here are simple and mainly descriptive, being based on the frequencies of word groupings in different collections of books digitised so far by Google. The free Google Books Ngram Viewer tool (http://books.google.com/ngrams) analyses frequencies of words, and phrases of a given length of n words (called *n-grams* or *ngrams*) for $n \leq 5$, amongst all words or phrases of the same length in Google's digital library. Normalisation through dividing by the number of all n-grams in all digitised books published in a given year is intended to ensure inter-temporal comparability of the results. More specific details on the tool and the methods are available in the paper by Michel et al. (2011). A parallel endeavour to the present work has been undertaken by Héran (2013), who focussed on the presentation of various entries in the 'demographic vocabulary'.

In our case we aim to go beyond a simple listing and additionally discuss some analytical possibilities offered by the Ngram tool. We first illustrate the approach on the example of the very name of the scientific discipline dealing with human populations, initially known as 'political arithmetick' thanks to William Petty's (1690) seminal work. As shown in Figure 1, 'demography' and 'population studies' are relatively new labels, both gaining in prominence only in the second half of the 20th century, with a clear dominance of the former.

Several comments need to be made with respect to the interpretation of the figures presented in this paper. Firstly, the lines present trends that have been manually smoothed by using five-term moving averages. Secondly, unless clearly stated otherwise, the queries have been limited in scope to English-language books, and are case-insensitive. Thirdly, the Ngram Viewer enables the combining (adding, subtracting, and dividing) of frequencies for different words and phrases, a technique which has been used in Figure 1 to allow for alternative spellings of the word 'arithmetic'.

In Figure 1, frequencies for 'demography' and 'population studies' are normalised by different *n*-gram counts, which explains some of the differences in magnitude. The apparent decline in the relative frequencies of these two terms does not signify a demise of the discipline, but quite the contrary: in terms of absolute numbers these terms have witnessed a near-exponential increase in prevalence in the Google books collection since the second half of the 20th century, but the overall number of different *n*-grams in digitized volumes has increased at an even greater pace (for a discussion of the increase of the information volume since Gutenberg, see the Introduction to Silver 2012).

Figure 1: Relative frequencies of different labels for the science of population in Google books

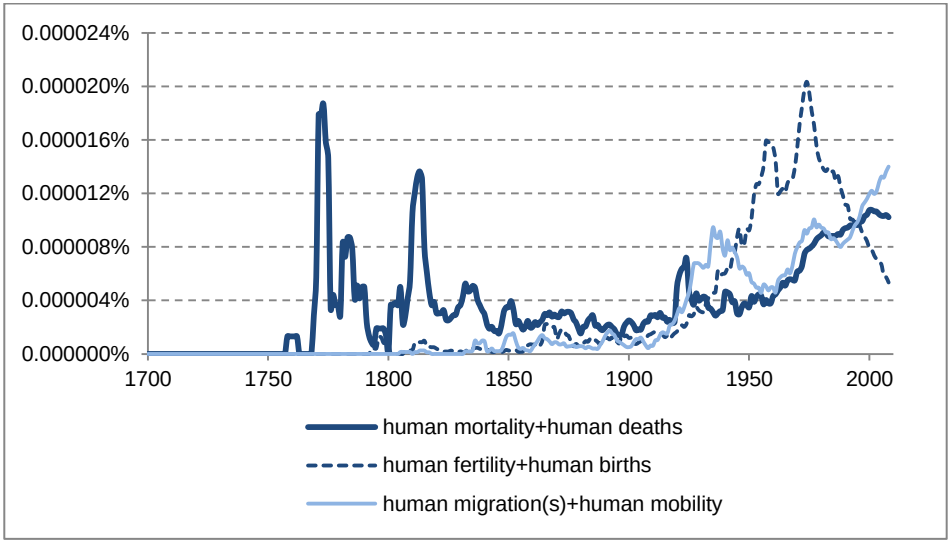


* Occurrences of 'demography' before its debut in 1855 (Courgeau 2012) are probably artefacts/scanning errors.
** Query included alternative spellings: 'political arithmetic', 'political arithmetick' and 'political arithmetics'.
Source: Google books Ngram Viewer, http://books.google.com/ngrams, English corpus, queried on 3.01.2014.

Our second example examines different components of population change: fertility, mortality, and migration. Here, in order to ensure that the search results are as closely related to demography as possible, we have decided to preface all queries with the adjective 'human'. We have also extended the searches to include 'births', 'deaths', and 'migrations', restricted to their plural grammatical form to obtain most of the matches from the scientific domain. The results should still be seen as approximate but, as illustrated in Figure 2, some trends in the relative importance of the three demographic parameters become apparent.

Unsurprisingly, mortality seemed of great importance in the "age of pestilence and famine" (see Omran, 1971), and also has been gaining prominence throughout the 20th century, when most of the modern gains in life expectancy took place. The relative frequency for fertility peaked in the mid-1970s, but for migration and mobility the trend is clearly upward, save for a temporary decline after the 1973 oil crisis. Also unsurprisingly, as the costs of migration decline, it turns into an ever-more important piece of the demographic balancing equation, which should not be ignored. Of course, the above-mentioned caveats on the interpretation of relative frequencies versus absolute numbers of *n*-grams remain in force.

Figure 2: Relative frequencies for different components of population change in Google books



Source: Google books Ngram Viewer, http://books.google.com/ngrams, English corpus, queried on 1.05.2013.

Our third example is illustrated in Figure 3, where we present different demographic paradigms – from period (cross-sectional) analysis, through cohort (longitudinal) analysis since the 1950s, event history analysis since the 1980s, followed by multilevel analysis (Courgeau and Franck 2007). The two panels of Figure 3 differ with respect to the language: the upper panel (Figure 3a) is based on the English corpus of Google books, and the lower one (Figure 3b) on the French one. Note, however, that due to possible multiple meanings of the English term 'period analysis' in different disciplines of science (mathematics, physics, economics…), only the trend for 'cross-sectional analysis' is shown. In addition, since the English term 'longitudinal analysis' in the social sciences has proliferated heavily outside demographic applications, in the example in Figure 3a it is shown separately.

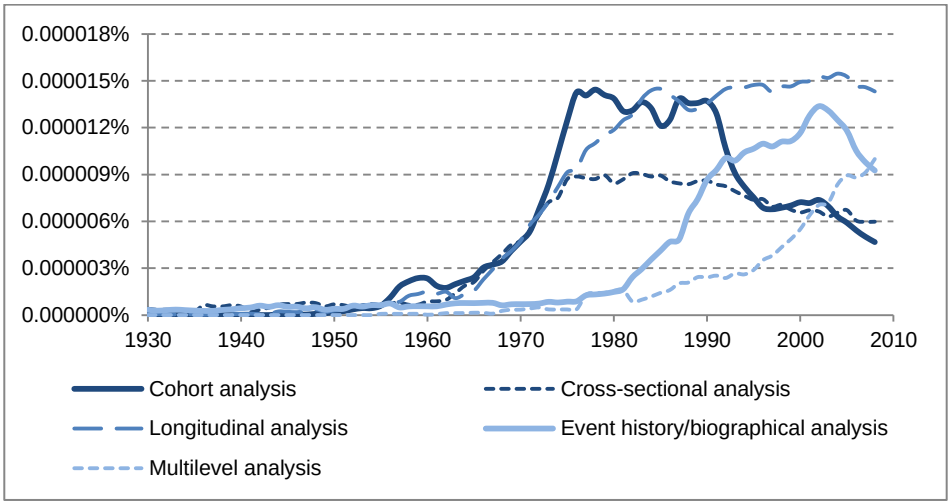
The trends observed in the 1960s for 'cross-sectional analysis' (Figure 3a), and for 'analyse transversale' (Figure 3b) have a clear interpretation: their appearance was necessitated by the emergence of cohort analysis, despite period analysis having been de facto used by demographers for many decades before. Hence, period/cross-sectional analysis as such did not emerge in the 1960s, but merely its label: previous to this there was only one way of performing demographic analyses.

The trends for 'event history/biographical analysis' and 'analyse biographique', despite different levels, largely exhibit similar directions in both languages, but indicate clear contamination with non-demographic meanings up until the 1970s – as noted by Courgeau (2012), the approach was introduced to demography only in the early 1980s. This is easy to verify by looking at the examples of results displayed by the Google Ngram tool alongside the trends: prior to the 1980s they mainly derive from such areas as psychology, literature, sociology, or aesthetics. Interestingly, since the 1980s until about 2000 there is a clear upward trend concerning 'event-history/biographical analysis' in both languages, which may owe to the role that French demographers played in the popularisation of the approach (idem).

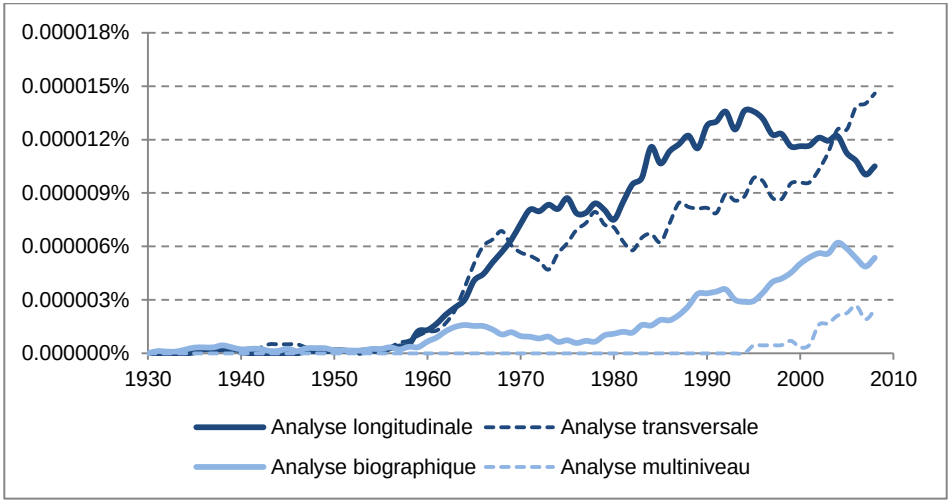
A comparison of the English and French graphs reveals some interesting properties, with the trends in French being more clearly marked. This calls for an interpretation of the emerging differences. Firstly, the underlying trends in the numbers of *n*-grams and their associated frequencies visibly differ between the French and English corpora, as illustrated in Figure 4 in the example of the terms 'demography' and 'démographie'. In other words, some of the differences may be due to variation in the normalisation constants applied.

Figure 3: Different paradigms related to population science since 1930 in Google books

a) English corpus



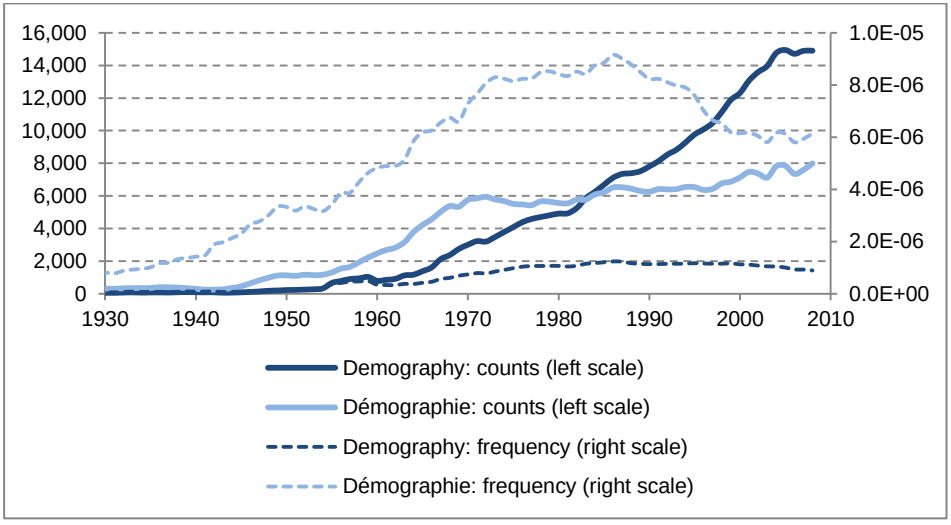
b) French corpus



Source: Google books Ngram Viewer, http://books.google.com/ngrams, different corpora, queried on 3.01.2014.

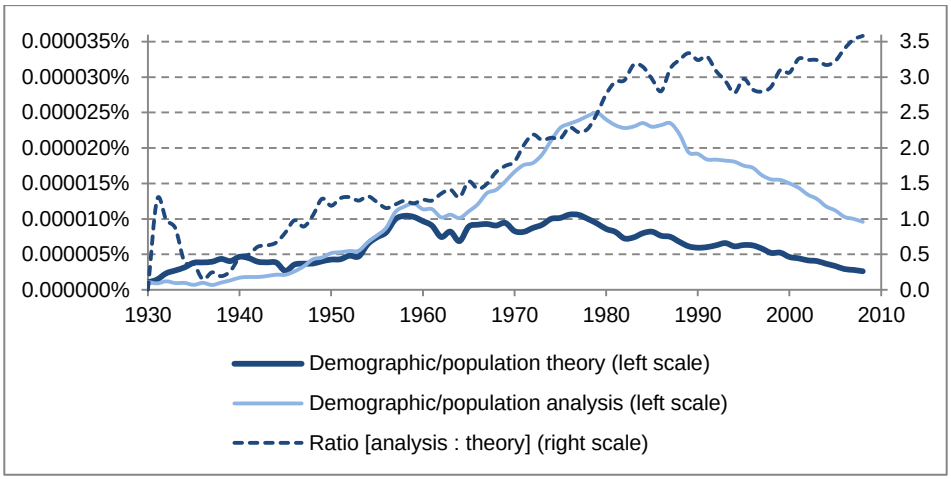
The final empirical example presented in this section is related to the relative importance of theoretical undertakings in demography, as compared with the analytical inclinations of the discipline. Figure 5 illustrates the respective trends: the former approximated by the sum of frequencies for 'demographic theory' and 'population theory', and the latter by 'demographic analysis' and 'population analysis'. The relative decrease in importance of demographic theorising, gradually occurring since the 1960s, is paramount. Even though the caveats about the disparity between numbers and frequencies remain in force, the ratio of the number of *n*-grams with 'analysis' to the ones with 'theory' is clearly increasing, as indicated by the dashed line in Figure 5. This supports the view of an important and growing gap in contemporary demography whereby well-developed analytical techniques are not backed up by theoretical foundations (e.g., Burch 2003).

Figure 4: Counts and frequencies of 'demography' and 'démographie' since 1930 in Google books



Source: Google books Ngram Viewer, http://books.google.com/ngrams, English/French corpus, queried on 3.01.2014.

Figure 5: Frequencies related to demographic theory and analysis since 1930 in Google books



Source: Google books Ngram Viewer, http://books.google.com/ngrams, English corpus, queried on 3.01.2014.

3. Discussion and future prospects

All the results presented in Section 2 point to the necessity of caution, and suggest that ideally such form of the content analysis should be conducted for more than one language corpus. Still, with this caveat in mind, we argue that a quantitative analysis of demographic paradigms, terms, and ideas similar to the one presented above can help the discipline enhance its self-knowledge. At such a general level, in this example for the English and French collections the findings seem to support the hypothesis of 'cumulativity' in population sciences (Courgeau 2012, see also Crimmins 1993 and Keyfitz 1993 for discussion), in which the new paradigms complement rather than substitute the existing ones.

There are many ways in which such an analysis could be extended: from analysing phrases from different language corpora (e.g., Chinese, Spanish, German, Russian, Italian); through looking at the prevalence of different demographic paradigms, theories and concepts as well as the interactions between them; to attempts at the prediction of future trends and the identification of 'hot topics' of demographic thought. Some additional ideas for analysing the *n*-gram output are offered by Michel et al. (2011).

In that respect, possible applications and extensions of the analysis presented in this paper include the detection of signals that could suggest changes to the methods or objects of demographic enquiry. Even though such shifts, by their very nature, are extremely difficult to identify ex ante, especially given current rapid developments in data collection and analytical methods, the proposed exploration can help assess the viability of some of these recent ideas. Besides, as noted by Crimmins (1993: 588), "formal demography is one area that has been characterized by continuity [based] on a long heritage, even while steady progress is made in the development of methodology and analytic techniques".

Personally, we believe that one area to keep an eye on is related to simulation modelling, including micro-simulation and other types of similar individual-level and multilevel approaches. In particular, an emerging paradigm here may be related to system-based modelling, which includes, for example, complex systems simulations and agent-based models (for pioneering work in demography, see Billari and Prskawetz 2003). Such methods not only allow bringing the context directly into the analysis, as in multilevel models (Crimmins 1993), but also have the potential to analyse the interactions between various systems comprised of individuals, groups, and institutions. In this way they can address some of the theoretical challenges of population sciences, mentioned e.g., by Xie (2000), Burch (2003), and Courgeau (2012), and also presented in Figure 5.

Of course, it is difficult to determine ex ante exactly what form such a new analytical paradigm would take. So far (as of September 2013) the Google Books collection of *n*-grams contains only a handful of occurrences of the phrase "Agent-Based Computational Demography" since 2003, so it is difficult to predict any lasting trend on that basis. However, the Google Books collection largely (although not entirely, as can be seen from sample results in French) omits information on relevant journal articles, in this case e.g., in *Demography*, *Demographic Research* and *Journal of Artificial Societies and Social Simulation*. In addition, some demographic books might have not been digitised by Google. Hence, to aid the 'early warning' process, analyses like this could be supplemented by more thorough bibliometric enquiries (cf. Teachman, Paasch, and Carver 1993; van Dalen and Henkens 2001), focusing on the usage of key words and phrases in different publications. The extent of the inclusion of journal articles in the Google books collection warrants a separate enquiry.

Furthermore, there are several important caveats that need to be made when conducting analyses based on *n*-grams. Most importantly, the query terms may be ambiguous. While 'demography' is used mainly in senses related to studies of human populations, other terms such as 'period analysis' are not, being shared with other areas of human knowledge. Future studies of empirical frequencies of *n*-grams for demographic applications thus needs to be based on a careful design of search queries,

cross-checked between different languages, in order to ensure as little ambiguity as possible. Ideally, the analysis should strive for one-to-one relationships between search terms and paradigms or approaches, indicating semantic unambiguity. In reality, there are examples of one-to-many (e.g., period analysis), many-to-one, and many-to-many (e.g., event-history analysis and biographical analysis) relationships. In such cases, even with well-devised queries, results are still very approximate.

Separate challenges involve the normalisation of Google Ngram output and the design of appropriate measures for presentation. The standard normalisation, performed through dividing by a total annual numbers of *n*-grams, may be found problematic, as it may artificially decrease the frequency of *n*-grams in the most recent years due to the constant inflow of new elements into the Google Books Ngram database (Bentley et al. 2012, Acerbi 2013). As an alternative, normalisation by the number of occurrences of definite articles ('the' in English) has been proposed (idem), although there are suggestions that it may lead to an opposite problem: artificial inflation of the most recent frequencies (Acerbi 2013). In any case, in the long run the normalisation of output needs attention.

Overall, however, the approach discussed in this paper is promising. Being a part of a wider area of quantitative content analysis, it remains open for further formal enquiries (for a recent overview of potentially applicable methods, see e.g., Krippendorff 2012). The existing examples of bibliographic studies in demography, from citation analysis (Keyfitz 1993; van Dalen and Henkens 2001) to enquiries of subject areas, characteristics of authors, and methods used (Teachman, Paasch, and Carver 1993) attest to the value of the approach. Amongst social science disciplines we think that demography, given its empirical slant, is a prime candidate for experimenting with what we see as a potentially very promising and fruitful method of philosophical-scientific investigation.

4. Challenge: Towards the history of population thought[5]

This research is by no means complete. Instead of simply concluding, we would like to open these ideas to discussion amongst the demographic community. In particular, we posit that demography is ripe for establishing a new sub-discipline, the History of Population Thought, possibly with its own dedicated journal. Similar endeavours exist in other fields, from science in general (with periodicals such as *Isis* or *Studies in History and Philosophy of Science*) to economics in particular, to name just one of the

[5] We thank the two anonymous Reviewers for their generous suggestions, many of which are included in this section.

social science disciplines related to demography (e.g., with *Journal of the History of Economic Thought*).

In the context of population sciences the focus of this sub-discipline could be on examining historical changes in various demographic paradigms, perspectives, theories, concepts, methods, models, and tools of analysis, ideally in a multilingual setting. Another experiment could be to chart a map of demography, either by looking at linkages and distances between concepts, or indeed between different authors, in a similar way as was done in the Literature Map project (Gibney n.d.), but based on citations rather than readers' preferences. In this way we hope that through the lens of such formal History of Population Thought the demographic community would gain more insight into our own discipline, and that this would facilitate a debate on the future of the population sciences in the 21st century.

5. Acknowledgements

A previous version of this paper was first presented at Chaire Quetelet, Louvain-la-Neuve, in November 2013. JB and ES gratefully acknowledge the Engineering and Physical Sciences Research Council grant EP/H021698/1 "Care Life Cycle", funded within the 'Complexity Science in the Real World' theme. The authors are grateful to two anonymous Reviewers and the Associate Editor of Demographic Research, Parfait Eloundou-Enyegue, for their comments that helped us improve the earlier draft. All views and interpretations in this paper are those of the authors and should not be attributed to any institution with which they are affiliated.

References

Acerbi, A. (2013). Normalization biases in Google Ngram [electronic resource]. acerbialberto.wordpress.com/2013/04/14/normalisation-biases-in-google-ngram

Bentley, R.A., Garnett, P., O'Brien, M.J., and Brock, W.A. (2012) Word Diffusion and Climate Science. *PLoS ONE* 7(11): e47966. doi:10.1371/journal.pone.0047966.

Billari, F. and Prskawetz, A. (eds.) (2003). *Agent-based computational demography. Using simulation to improve our understanding of demographic behaviour*. Heidelberg, New York: Physica-Verlag. doi:10.1007/978-3-7908-2715-6.

Burch, T. (2003). Demography in a new key: A theory of population theory. *Demographic Research* 9(11): 263–284. doi:10.4054/DemRes.2003.9.11.

Courgeau, D. (2012). *Probability and Social Science. Methodological Relationships between the two Approaches*. Dordrecht: Springer. doi:10.1007/978-94-007-2879-0.

Courgeau, D. and Franck, R. (2007). Demography, a fully formed science or a science in the making. *Population–E* 62(1): 39–45. doi:10.3917/pope.701.0039.

Crimmins, E.M. (1993). Demography: The Past 30 Years, the Present, and the Future. *Demography* 30(4): 579–591. doi:10.2307/2061807.

Gibney, M. (n.d.). Literature Map [electronic resource]. www.literature-map.com.

Granger, G.-G. (1994). *Formes, opérations, objets*. Paris : Librairie Philosophique Vrin.

Héran, F. (2013). Demography and its vocabulary over the centuries: A digital exploration. *Population & Societies* 503(November 2013).

Keyfitz, N. (1993). Thirty Years of Demography and «Demography». *Demography* 30(4): 533–549. doi:10.2307/2061805.

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. London: Sage.

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Brockman, W., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331: 176–182. doi:10.1126/science.1199644.

Morgan, S.P. and Lynch, S.M. (2001). Success and Future of Demography. The Role of Data and Methods. *Annals of the New York Academy of Sciences* 954: 35–51. doi:10.1111/j.1749-6632.2001.tb02745.x.

Omran, A. (1971). The Epidemiologic Transition: A Theory of the Epidemiology of Population Change. *The Milbank Memorial Fund Quarterly* 49(4): 509–538. doi:10.2307/3349375.

Petty, W. (1690). *Political Arithmetick*. London: Robert Clavel & Hen. Mortlock at St Paul's Churchyard.

Silver, N. (2012) *The Signal and the Noise. The Art and Science of Prediction*. London: Penguin.

Teachman, J.D., Paasch, K., and Carver, K.P. (1993). Thirty years of «Demography». *Demography* 30(4): 523–532. doi:10.2307/2061804.

van Dalen, H.P. and Henkens, K. (2001) What makes a scientific article influential? The case of demographers. *Scientometrics* 50(3): 455–482. doi:10.1023/A:1010510831718.

Xie, Y. (2000). Demography: Past, Present and Future. *Journal of the American Statistical Association* 95(450): 670–673. doi:10.1080/01621459.2000.10474248.