Quantifying paradigm change in demography

Jakub Bijak[1]✉, Daniel Courgeau[2], Eric Silverman[1], and Robert Franck[3]

[1] Social Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom

[2] Research Director Emeritus, Institut national d'études démographiques (INED), Paris, France

[3] Professor Emeritus, Université catholique de Louvain, Louvain-la-Neuve, Belgium

✉ Corresponding author, email: j.bijak@soton.ac.uk

Abstract

Demography is a uniquely empirical research area amongst 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. The paper aims to include measurable data in the studies of changes in selected demographic paradigms and theories. 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. The tentative findings corroborate the shifts in the demographic paradigms identified in the literature – from cross-sectional, through longitudinal, to event-history and multilevel perspectives. The findings identify a promising area of enquiry into the development of demography as a social science discipline. Still, more collaborative work in this area is needed, and we actively solicit an interactive discussion on quantifying the changes in demographic paradigms and theories.

Key words

Demographic paradigms – Empiricism – Google books – History of demography – N-grams

* * * * *

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 2008; 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 overall demographic nomenclature and to the studies of different components of demographic change. Section 3 contains a discussion of some of the findings, followed by a brief evaluation of some of the potential benefits and limitations of the application of the proposed method to the study of theoretical and paradigmatic change in demography. We conclude by proposing an open challenge for the demographic community in Section 4.

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 axioms (idem). Such developments could ultimately facilitate theory-building – something that is seen as one of the key challenges of 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 this aim.

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 the research results in the public policy field (for a discussion, see e.g. Xie 2000, and Morgan and Lynch 2008). In addition, demographic works are also on average more frequently cited 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 yet, 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.

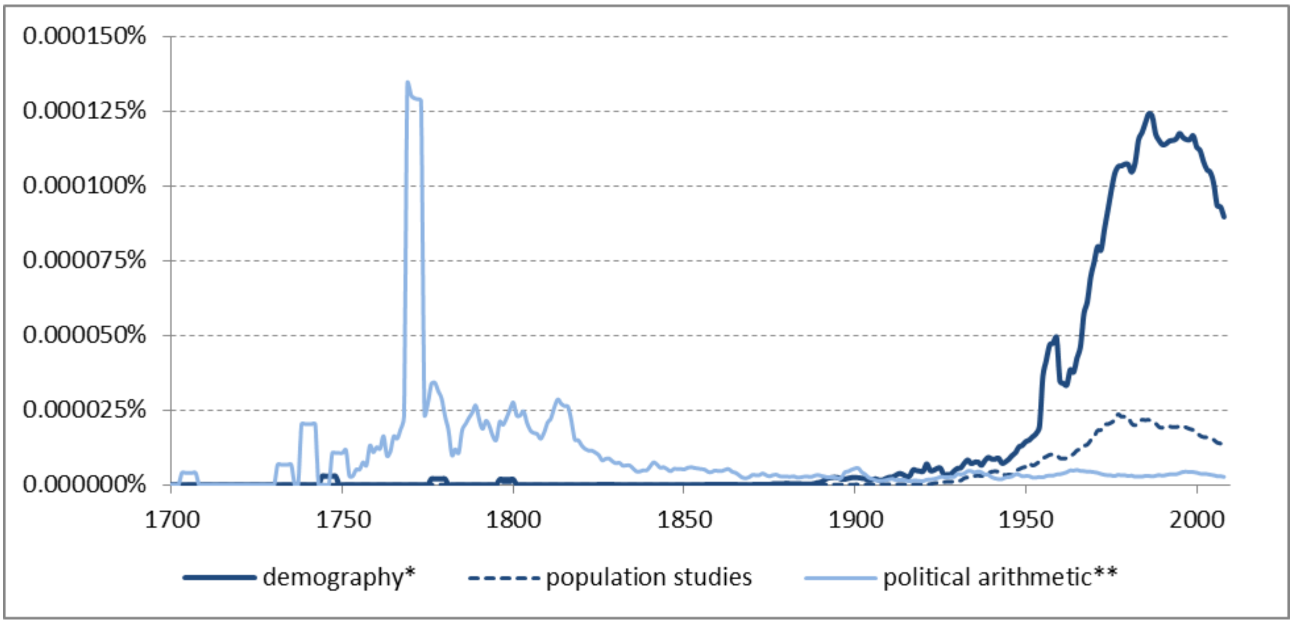
The examples presented in this paper are very simple and purely descriptive, being based on the frequencies of word groupings in different collections of books digitised so far by Google. The free *Google books Ngram Viewer* tool (http://books.google.com/ngrams) analyses frequencies of words, and phrases 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).

In our case, 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.

Five comments need to be made with respect to the interpretation of Figure 1. Firstly, the lines present trends that have been smoothed by using five-term moving averages. Secondly, the scope of the query has been limited to English-language books. Thirdly, the Ngram Viewer enables combining (adding, subtracting and dividing) frequencies for different words and phrases, which has been used here to allow for alternative spellings of the word 'arithmetic'. Fourthly, frequencies for 'demography' and 'population studies' are normalised by different *n*-gram counts, which explains some of the differences in magnitude. Finally, 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 *Introduction* to Silver 2012).

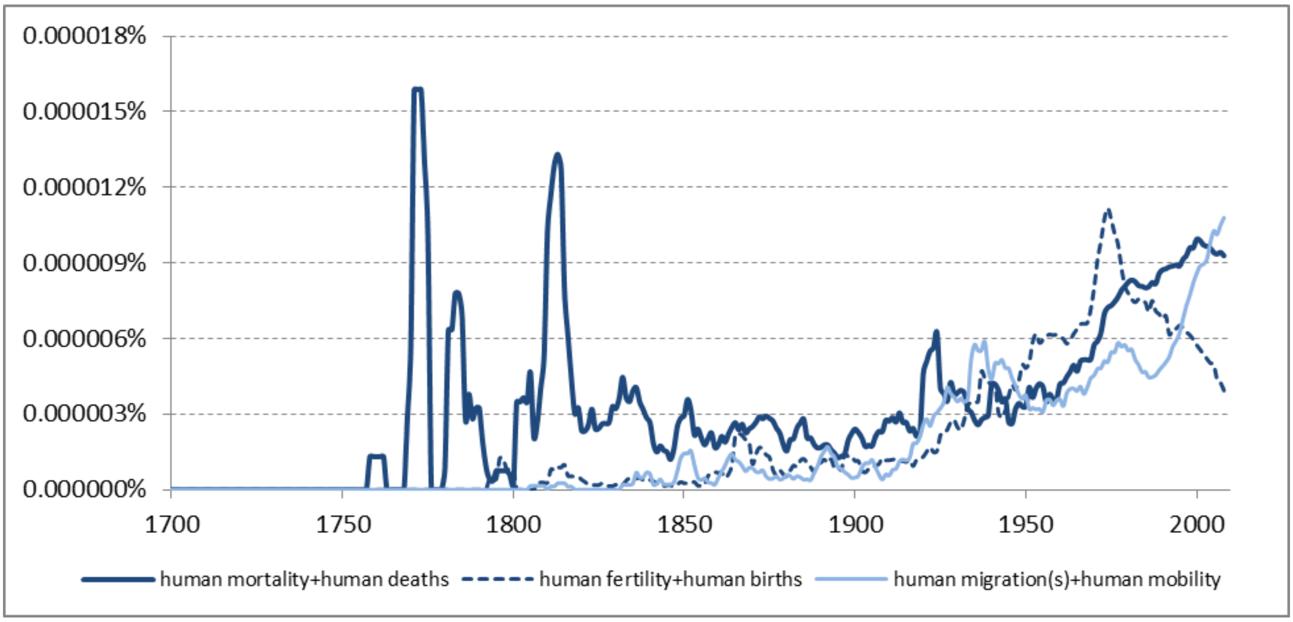
The second example refers to different components of population change: fertility, mortality and migration. Here in order to ensure that the search results are as closely related to demography as possible, we have decided to precede all queries by the adjective 'human'. We have also extended the searches to include 'births', 'deaths' and 'migrations', restricted to their plural grammatical form to get 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.

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 1.05.2013

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

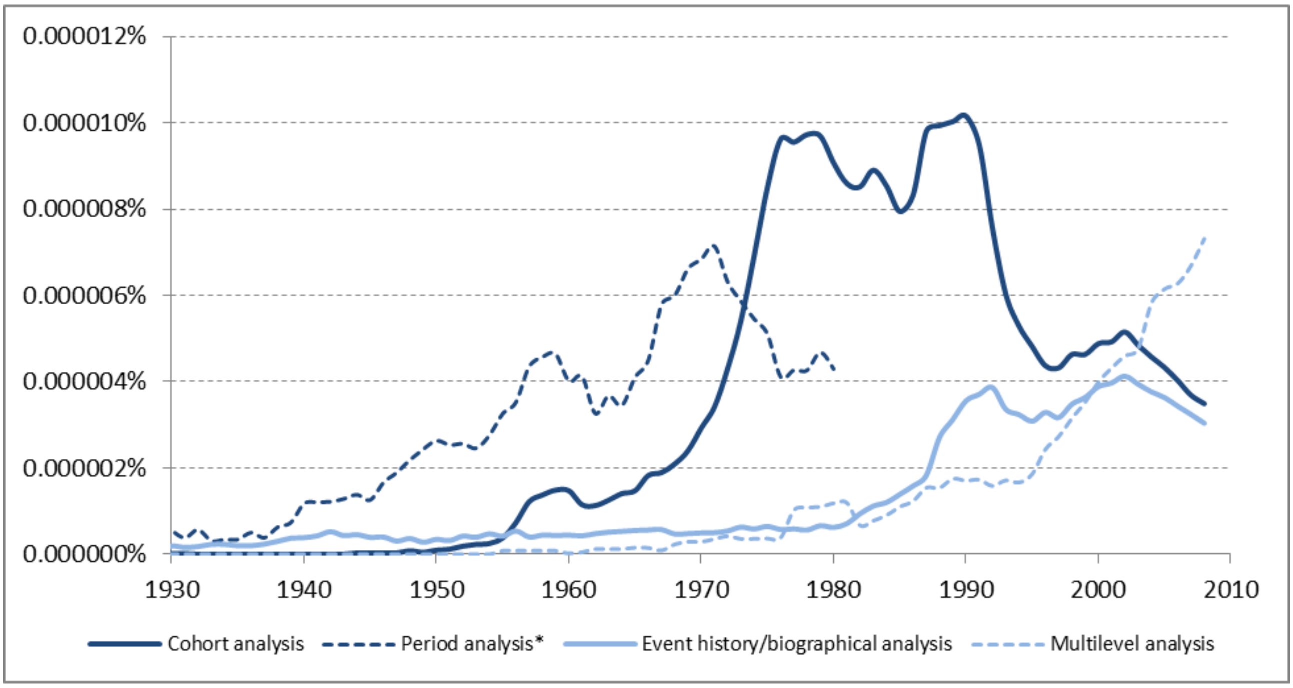


Source: Google books Ngram Viewer, http://books.google.com/ngrams, English corpus, queried on 1.05.2013

Mortality seemed of great importance in the ‘age of pestilence and famine’ (see Omran, 1971), but has also been gaining prominence throughout the 20th century, when most of the modern gains in life expectancy took place. The relative frequency for fertility has peaked in mid-1970s, but for migration and mobility, save for a temporary decline after the 1973 oil crisis, the trend is clearly upwards. This suggests that migration is becoming an ever-more important piece of the demographic balance 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.

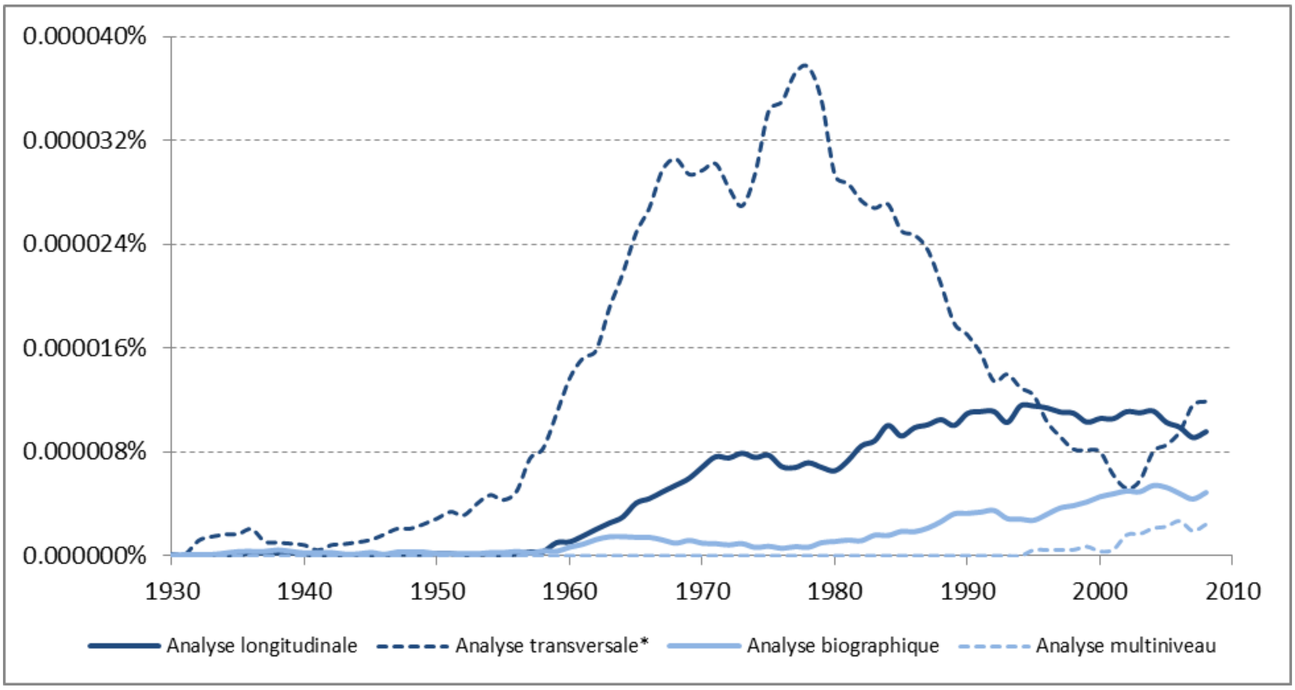
The third example is illustrated in Figures 3 and 4, the difference between which is in the language: the former is based on the English corpus of Google books, and the latter on French. In these figures we present different demographic paradigms – from period analysis, through cohort analysis since the 1950s, event history analysis since the 1980s, followed by multilevel analysis (Courgeau and Franck 2007). Notably, due to possible multiple meanings of the English term 'period analysis' in different disciplines of science, the trend for this term as shown in Figure 3 is approximated by the difference between 'demographic analysis + population analysis' and 'cohort analysis'. Due to the multiple paradigms present in demography in recent decades, this approximation is shown only until 1980. In addition, since the term 'longitudinal analysis' has proliferated heavily outside demographic applications, in the example in Figure 3 only the reference to 'cohort analysis' has been retained.

Figure 3. Different paradigms related to population science since 1930 in Google books (English)



* Approximated by differences between 'demographic analysis + population analysis' and 'cohort analysis'
Source: Google books Ngram Viewer, http://books.google.com/ngrams, English corpus, queried on 3.05.2013

Figure 4. Different paradigms related to population science since 1930 in Google books (French)



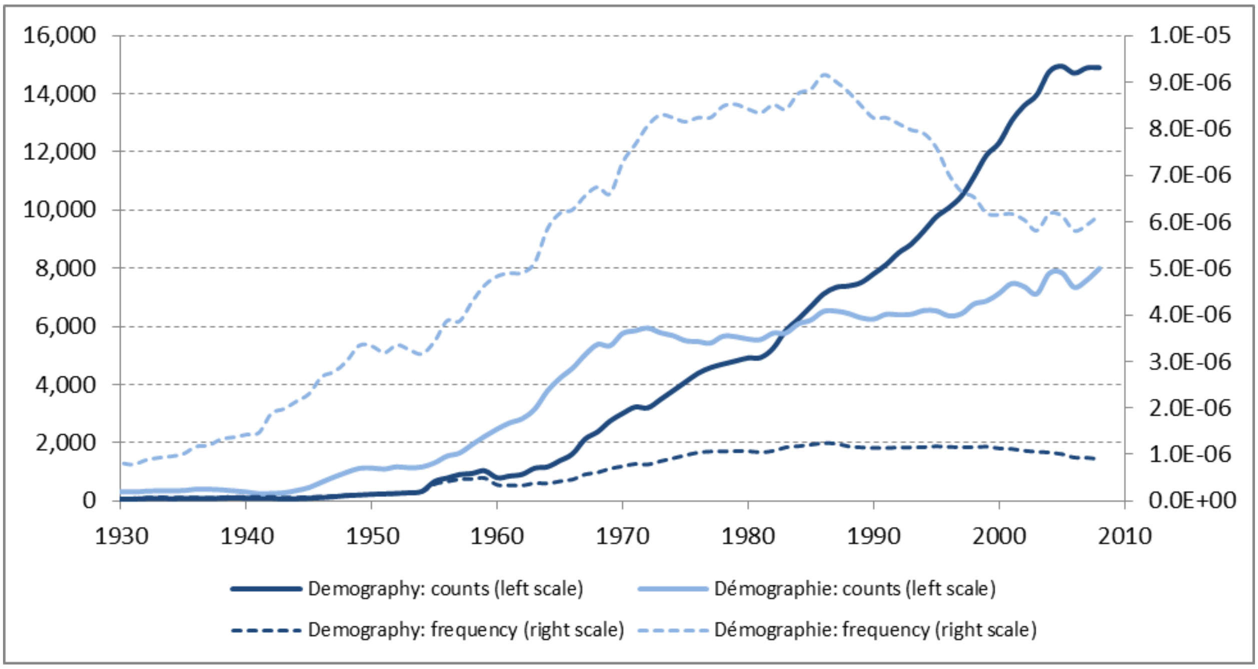
* Approximated by differences between 'analyse démographique + analyse transversale' and 'analyse longitudinale + analyse biographique + analyse multiniveau', with 'analyse biographique' included only after 1980 (see text).
Source: Google books Ngram Viewer, http://books.google.com/ngrams, French corpus, queried on 5.10.2013

A comparison of the English and French graphs reveals some interesting properties, with the trends in French being much 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 5 in the example of the terms 'demography' and 'démographie'. In other words, some of the differences may be due to variation in the normalisation procedures applied.

Secondly, as seen through much more steady trends, some of the French terms, such as 'analyse longitudinale',seem more unambiguous than their English counterparts, and are less likely to be conflated with similar expressions outside of demography, or even social sciences. The ballpark trends for 'analyse transversale' in French and 'period analysis' in English also differ substantially, especially in terms of scale – possibly due to different approximations used, or different normalising constants in the denominator (see also Figure 5).

The trends for 'event history', 'biographical analysis' and 'analyse biographique' exhibit similar levels 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 together with the trends: prior to 1980s they mainly consider such areas like psychology, literature, sociology, or aesthetics. Interestingly, since the 1980s trends concerning the event-history/biographical analysis in both languages are relatively comparable, which may be owing to the role that French demographers played in the popularisation of the approach (*idem*).

Figure 5. 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 5.09.2013

Finally, the trend for non-approximated 'analyse transversale' alone (not shown in Figure 4, but parallel to 'analyse longitudinale') has a clear interpretation: its 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 analysis as such did not emerge in the 1960s, but merely its label: before this there was only one way of performing 'analyse démographique'.

3. Discussion and Future Prospects

All the results presented in Section 2 point to the necessity of caution, and suggest that ideally the analysis should be conducted for more than one language corpus. When examined 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), in which the new paradigms *complement* rather than *substitute* the existing ones. Still, with these caveats 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.

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 further changes to the methods of demographic enquiry. Personally, we believe that agent-based modelling (e.g. Billari and Prskawetz, 2003) may be one such new paradigm, as it is potentially capable of addressing some of the theoretical challenges of population sciences, mentioned e.g. by Xie (2000), Burch (2003) and Courgeau (2012). 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. However, this source 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, such an analysis could be this supplemented by more thorough bibliometric enquiries (cf. van Dalen and Henkens 2001), focusing on the usage of particular key words and phrases in different demographic publications. The extent of the inclusion of journal articles in the Google books collection warrants a separate enquiry.

On the other hand, there are some important caveats that need to be kept in mind when conducting such analyses based on *n*-grams. Most importantly, the terms used 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 will thus need to be based on the careful design of search queries, and cross-checked between different languages, in order to ensure as little ambiguity as possible. However, even with well-devised queries, results should still be characterised as 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. Dubbed "culturomics" by its original creators, this methodology is envisioned to extend "the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities" (Michel et al. 2011: 176). 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 enquiry.

4. Challenge

This research is by no means complete. Instead of simply concluding, we would like to open these ideas to discussion amongst the demographic community. We would welcome any feedback notes and challenges, to be communicated directly to the authors[*]. In particular, we would appreciate further suggestions on measurement issues, as well as on which trends to examine. The latter could involve various demographic paradigms, perspectives, theories, concepts, methods, models and tools of analysis. The ultimate aim of such an interactive experiment would be to chart a conceptual and paradigmatic map of demography, co-authored by all the contributors to the discussion. In this way we hope that the demographic community would gain more insight into our own discipline, and that such an exercise would facilitate a debate on the future of the population sciences in the 21st century.

Acknowledgements

JB and ES gratefully acknowledge the Engineering and Physical Sciences Research Council (EPSRC) grant EP/H021698/1 "Care Life Cycle", funded within the 'Complexity Science in the Real World' theme. All the views and interpretations reported in this paper are those of the authors and should not be attributed to any institution with which they are affiliated.

[*] A journal submission based on this paper is currently pending – if it gets published, we will naturally also very much welcome formal letters through the response facility of the journal.

References

Acerbi, A. (2013) Normalization biases in Google Ngram. Blog entry, 14 March 2013 [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. www.plosone.org/article/info: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.

Burch, T. (2003). Demography in a new key: A theory of population theory. *Demographic Research*, 9(11), 263–284. www.demographic-research.org/Volumes/Vol9/11

Courgeau, D. (2012). *Probability and Social Science. Methodological Relationships between the two Approaches*. "Methodos" Series, vol. 10. Dordrecht: Springer.

Courgeau, D. and Franck, R. (2007). Demography, a fully formed science or a science in the making, *Population-E*, 62(1), 39-45.

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

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.

Morgan, S.P., and Lynch, S.M. (2008). Success and Future of Demography. The Role of Data and Methods. *Annals of the New York Academy of Sciences*, 954, 35–51.

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

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.

van Dalen, H.P., and Henkens, K. (2001) What makes a scientific article influential? The case of demographers. *Scientometrics*, 50(3), 455-482.

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

■