Local Population Studies in the Era of 'Big Data'

Andrew Hinde

Abstract

Recent years have seen a proliferation of very large data sets in historical demography, many of which have assembled individual-level data for entire populations. It might be thought that these data sets pose a challenge to local population studies, as they remove one of the rationales for work at the local level: that research using individual-level data was only logistically possible for small populations. This paper argues that, to the contrary, the advent of 'big data' sets provides new opportunities for work on local populations. Many of the old arguments in favour of local history still hold, but 'big data' can direct researchers to those places where local studies can potentially make the biggest contribution, and thus give local population studies a new lease of life.

Introduction

Recent years have seen a proliferation of very large data sets in historical demography, many of which have assembled individual-level data for entire populations. It might be thought that these data sets pose a challenge to local population studies, as they remove one of the rationales for work at the local level: that research using individual-level data was only logistically possible for small populations. This paper argues that, to the contrary, the advent of 'big data' sets provides new opportunities for work on local populations. Many of the old arguments in favour of local history still hold, but 'big data' can direct researchers to those places where local studies can potentially make the biggest contribution, and thus give local population studies a new lease of life.

The paper begins by discussing what the term 'big data' might mean in the conext of historical demography. It then discusses (and rejects) the argument that 'big data' have superseded the need for local studies, but acknowledges that other challenges, based around the institutionalisation of research in the academic world, remain. The next sections try to formulate the relationship between analyses based on 'big data' and local studies. The argument intersects with the idea of a 'sense of place', which has often been invoked in discussions of local history.

'Big data' in historical demography

There is no universally accepted definition of 'big data'. Most definitions make the point that the data are too large to be analysed effectively by standard software.[1] But this is a rather slippery description, as the capacity of 'standard software' is increasing with time, so that the size at which a data set might meet this definition is increasing too. From the perspective of historical demography, it may be better to define 'big data' in relation to specific data sets that have been recently put together, or that are still under construction.

So, for example, there is the Integrated Census Microdata (I-CeM) project for England and Wales, which has made available (at least to historians with an institutional affiliation) individual-level data for the entire population of those countries for the censuses of 1851, 1861, 1881, 1891, 1901 and 1911, a total of more than 183 million records.[2] Further afield, there is the Integrated Public Use Microdata Series (IPUMS) based at the University of Minnesota, which includes census data from around the world from the eighteenth century to the present day.[3] Outside academia, vast collections of data are held by commercial bodies in the field of family history, such as Ancestry and FindMyPast.[4] Some of these data sets are gigantic (IPUMS historical census series for the United States includes over one *billion* records).[5]

Arguably. historical demography was the social science that motivated the first collection of 'big data', starting with the sickness surveys of the nineteenth century. These culminated in Alfred Watson's study of sickness and mortality among the Independent Order of Odd Fellows between 1893 and 1897, which included more than seven million weeks of sickness and almost three million years of life exposed to the risk of falling sick.[6] Indeed, the journal *Local Population Studies* and the Local Population Studies Society were founded as a consequence of an exercise in constructing a 'big data' set in an era when the labour of many people was required in order to transcribe and input such a volume of data. The resulting data set consists of monthly totals of baptisms, marriages and burials for a sample

[1] See, for example, C. Snijders, U. Matzat and U.-D. Reips, ' "Big data": big gaps of knowledge in the field of internet science', *International Journal of Internet Science*, 7 (2012), pp. 1–5, here at p. 1. http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf [accessed 4 June 2019].

[2] K. Schürer, and E. Higgs, E. *Integrated Census Microdata (I-CeM), 1851–1911* [data collection] Colchester, England: UK Data Archive [distributor], 2014. SN 7481; E. Higgs, C. Jones, K. Schürer and A. Wilkinson, *Integrated Census Microdata (I-CeM) Guide* (Colchester, 2013). https://www1.essex.ac.uk/history/research/icem/documents/icem-guide-version-2-2015.pdf [accessed 4 June 2019]. Most (though not all) of the I-CeM data are publicly available from the United Kingdom Data Archive. Anyone with an institutional affiliation can access these data by visiting the web page https://icem.data-archive.ac.uk/#step1 [accessed 22 June 2019]. The data that are not available are mainly individuals' names. Even these can be accessed by researchers by application to the I-CeM team.

[3] See https://ipums.org/what-is-ipums [accessed 4 June 2019].

[4] https://www.ancestry.co.uk/ [accessed 4 June 2019]; https://www.findmypast.co.uk/ [accessed 4 June 2019].

[5] For an example of a paper using just a fraction of one of these data sets, see N. Cummins, M. Kelly and C. Ó Gráda, 'Living standards and plague in London, 1560–1665', *Economic History Review*, 69 (2016), pp. 3–34. http://dx.doi.org/10.1111/ehr.12098. Cummins and his colleagues use some 920,000 burial records and 630,000 baptism records obtained from Ancestry to trace the geography of plague in sixteenth and seventeenth century London.

[6] A.W. Watson, *An Account of an Investigation of the Sickness and Mortality Experience of the I.O.O.F. Manchester Unity: during the Five Years 1893–1897* (Manchester, 1903), p. 21. Still earlier, of course, the London Bills of Mortality can be seen as a 'big data' set.

of 404 English parishes between 1538 and 1837, a total of around three million data items.[7] Other large data sets dealing with the historical demography of England and Wales are the data on deaths by age and cause compiled by Robert Woods and Nicola Shelton in the 600 or so registration districts for each of five decades between 1851 and 1900.[8]

When *Local Population Studies* was founded in 1968, there was a plan to use the data collected and submitted by the many local historians who contributed to the 404-parish data set to promote the study of local population change. The primary purpose of the data collection exercise, though, was not local at all, but to provide a sufficiently large and representative sample of vital events for the period between 1538 and the nineteenth century to allow the population of England *as a whole* to be reconstructed. Only a limited discussion of the variety of local experience featured in the book which formed the principal output of the project.[9] It was hoped that the project would encourage the contributors to conduct their own research on the parishes and localities upon which they worked and, for a few individual localities, this happened.[10] But the vast majority of the 404 parishes for which data were collated did not feature in these local studies. Much more local and regional work could be done using the data.[11]

Challenges for local population studies

In years past, local research was necessary to make progress on many questions of interest to historical demographers. There were several reasons for this. First, some source materials were inherently local, not only in their content but in their physical location (the parish

[7] Assuming that the 'average' parish has 200 years' worth of monthly data for the three series of baptisms, marriages and burials, this produces 404 x 200 x 12 x 3 = 2,908,800 items of data. These data were subsequently complemented by the data sets assembled by the Cambridge Group for the History of Population and Social Structure using family reconstitution: see E.A. Wrigley, R.S. Davies, J.E. Oeppen and R.S. Schofield, *English Population History from Family Reconstitution 1580–1837* (Cambridge, 1997).

[8] This data set has close to two million independent items. See R. Woods, *Causes of Death in England and Wales, 1851–1860 to 1891–1900: the Decennial Supplements* [data collection]. Colchester, England: UK Data Archive [distributor], 2014. SN 3552. http://doi.org/10.5255/UKDA-SN-3552-1.

[9] E.A. Wrigley and R.S. Schofield, *The Population History of England, 1541–1871: a Reconstruction* (Cambridge, 1989), p. 39; the topic most thoroughly considered at the local level is mortality crises on pp. 645–96.

[10] The most thoroughly studied parish is Colyton in Devon: see E.A. Wrigley, 'Family limitation in pre-industrial England', *Economic History Review*, 19 (1966), pp. 82–109; E.A. Wrigley, 'The changing occupational structure of Colyton over two centuries', *Local Population Studies*, 18 (1977), pp. 9–21, https://doi.org/10.1111/j.1468-0289.1966.tb00962.x; P. Sharpe, 'Literally spinsters: a new interpretation of local economy and demography in Colyton in the seventeenth and eighteenth centuries', *Economic History Review*, 44 (1991), pp. 46–65, https://doi.org/10.1111/j.1468-0289.1991.tb01264.x. Other examples are Shepshed in Leicestershire: see D. Levene, *Family Formation in an Age of Nascent Capitalism* (London, 1977). For eight of the 404 parishes in Gloucestershire, see Janet Hudson, 'The incorporation of evidence about local nonconformity into parish population reconstruction', *Local Population Studies*, 80 (2008), pp. 39–58.

[11] It is partly to promote such work that the Local Population Studies Society is shortly to launch a project to make available online not only the data for the 404 parishes but for the many other parishes for which good quality data have been transcribed and checked in the years since the publication of Wrigley and Schofield, *Population History of England.*

registers of England, for example), so the construction of any consolidated database from these sources required local work. Second, it was considered that aggregate-level data were unable to provide satisfactory answers to many of these questions, especially (but not exclusively) those about the behaviour of individuals. Aggregate-level data were subject to problems such as the ecological fallacy, and changes and inconsistencies in the geographical limits of the aggregated populations.[12] However, the collection and processing of individual-level data was so laborious that only small populations could be studied. Moreover, sampling was often ruled out in cases where multi-source record linkage was to be used. This meant that individual researchers could hardly tackle data sets larger than about 60,000 cases.[13] Even where sampling was possible, teams of researchers were restricted to relatively modest samples, such as the 2 per cent sample of the 1851 census transcribed by a team led by Michael Anderson.[14]

Changing technology and the arrival of 'big data' have done away with these practical reasons for local population studies. Automatic character recognition, and the accumulated transcription work of countless family and local historians, mean that it is now possible for one or two individuals to process millions of cases of individual-level data. For example, what was considered in the 1980s to be a large study of household structure in the second half of the nineteenth century, analysed four small regions of England for the four censuses of 1851, 1861, 1871 and 1881 and used a total population of around 50,000 cases in the census years combined.[15] In 2018 a study of household structure was published using individual-level data for the entire population of England and Wales for the censuses of 1851 through to 1911 (apart from 1871).[16]

'Big data' poses other challenges to local studies. Academic historians' careers depend on attracting research grants and producing publications which are highly ranked by the Research Evaluation Framework. Success in this is much easier to achieve with large, national (or preferably international) projects. 'Big data' makes such projects possible even using individual-level data. International conferences of historical demography in recent

[12] The ecological fallacy is the name given to the fact that a correlation may be observed between two variables at the aggregate level (for example when using county data) when no correlation between the same two variables exists at the individual level.

[13] One of the largest datasets compiled in this era was the individual-level census data for the town of Keighley in West Yorkshire studied by Eilidh Garrett: see, for example, E.M. Garrett, 'The trials of labour: motherhood and employment in a nineteenth-century textile centre', *Continuity and Change*, 5 (1990), pp. 121–54, https://doi.org/10.1017/S0268416000003908.

[14] Anderson's sample was the largest data set compiled from the census enumerators' books until the transcription of the entire 1881 census made available through the Church of Jesus Christ of Latter-day Saints. See M. Anderson, 'Households, families and individuals: some preliminary results from the national sample from the 1851 census of Great Britain', *Continuity and Change*, 3 (1988), pp. 421–38, https://doi.org/10.1017/S0268416000004306.

[15] P.R.A. Hinde, 'Household structure, marriage and the institution of service in nineteenth century rural England', *Local Population Studies*, 35 (1985), pp. 43–51.

[16] K. Schürer, E.M. Garrett, H. Jaadla and A. Reid, 'Household and family structure in England and Wales (1851–1911): continuities and change', *Continuity and Change*, 33 (2018), pp. 365–411, https://doi.org/10.1017/S0268416018000243.

years invariably include sessions dealing with new data bases, or methodological topics related to the construction of such 'big data'.[17]

Alongside our ability to create 'big data' sets has been an increase in the variety of techniques designed to uncover patterns in such data. Approaches such as data mining and various methods of statistical learning can be applied to 'big data' and used to reveal regularities and patterns which are not obvious on an initial examination.[18] Although such methods have not hitherto been applied to historical demography data sets, this is likely to happen in the future and such work will probably appeal to journal editors and those awarding research grants.[19]

So why do local population studies?

Given these challenges: the attraction of large-scale, high-profile projects for academics, and the fact that individual-level data sets are no longer restricted to local populations for practical reasons, one might ask why continue do local population studies at all?

The first point to make in answer to this question is that the value of local history has been debated for many years, and many of the justifications for local studies given in the past still apply. The first two of these are well known and will be discussed only briefly. They are, first, that local studies provide a richness of context that studies of large populations rarely can.[20] It seems unlikely that the appearance of 'big data' is going to change this. Much 'big data' in historical demography includes limited numbers of variables, most of which are quantitative. It can therefore only provide a skeletal contextual picture.[21] Second, local studies offer an opportunity to study people's lives at a scale which reflects the scale on which those lives were lived and the range of territory with which people had connections. This territory may not be restricted to single communities, or

[17] For example, the European Society for Historical Demography conference in June 2019 included sessions entitled 'New Nordic databases: a momentum to historical demography in Nordic countries', 'Popular genealogy as citizen science' and 'Automating source transcription'.

[18] For an introduction to these methods, see for example, G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R* (New York, 2013), https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf [accessed 4 June 2019].

[19] This will be largely on the grounds of novelty, rather than an objective assessment of the likely contribution of such approaches. Past experience in other disciplines suggests that these approaches are new, but not as new as many believe; they will also potentially provide new insights, but will deliver less than is claimed when they are first introduced.

[20] These and other reasons for doing local history are set out in J.D.Marshall*, The Tyranny of the Discrete: a Discussion of the Problems of Local History in England* (Aldershot, 1997).

[21] Those major databases that have overcome this limitation have done so by restricting themselves to regional or local populations, and thereby becoming local or regional databases. An example is the Demographic Data Base (DDB) at the Center for Ageing and Demographic Research at Umeå University. For a discussion of how the DDB was set up, see S. Edvinsson, 'The Demographic Data Base at Umeå University: a resource for historical studies', in P.K. Hall, R. McCaa and G. Thorvaldsen (eds), *Handbook of International Historical Microdata for Population Research* (Minneapolis, 2000). https://international.ipums.org/international/microdata_handbook.shtml [accessed 4 June 2019]. More information about the Umeå University database is at https://www.umu.se/en/centre-for-demographic-and-ageing-research/databases/ [accessed 4 June 2019].

bounded by administrative units, but for most people, most of the time, social and economic interaction was limited geographically. This does not mean that these connections were always with places physically nearby, or that the strength of interactions was inversely related in a simple way to geographical distance, but merely to note that some connections were geographically more important than others for particular places.[22] Here, 'big data' can in some cases actually help delineate the shape of a locality's connectedness with the rest of the world and thereby help direct local studies to the appropriate target populations. The I-CeM census data can be used, for example, to map the birthplaces of the inhabitants of specific places in a given census, or to map the places of residence at any given census of those born in particular localities, and thus to delineate lifetime migration links both to and from specific localities.[23]

A third argument is more complex, but may help to provide a clearer exposition of the role of local studies in the future. We might start with the concept of *place* discussed by J.D. Marshall in his 1997 book, *The Tyranny of the Discrete*. Marshall says that place is 'not simply a matter of space … [but] … of human perception and recognition'.[24] It is associated with geographical location, both in relation to the physical environment and relative to other places.[25] It is also linked to the region over which everyday lives were lived. But partly, it consists of a 'sense of place', or 'a set of subjective views of place derived from residence within it'.[26] Unfortunately, as Marshall indicates, these 'concepts have not been readily adopted by historians, perhaps because they seem to point to some almost insoluble methodological problems', and he is unable to reach any conclusions as to what historians should do about them, save to reiterate a point well made by Dennis Mills that the geographical extent of local studies should not be circumscribed by 'arbitrarily defined administrative entities'.[27]

If Marshall is right, then it may be unprofitable to spend a lot of time trying to uncover the essence of 'place'. However, I argue that what matters for population studies is not so much what place consists in, but the extent to which, and the mechanisms through which,

[22] See M. Hardy, 'The Newfoundland trade and Devonian migration c. 1600–1850', *Local Population Studies*, 89 (2012), pp. 31–53, in which Hardy shows that because of historical trade links, Newfoundland had become, in some senses, much closer to Devon than were other parts of England.

[23] For an example of the kind of analysis that can be achieved, see K. Schürer and J. Day, 'Migration to London and the development of the north–south divide, 1851–1911', *Social History*, 44 (2019), pp. 26–56, https://doi.org/10.1080/03071022.2019.1545361. The potential of this kind of analysis is enormous. Twenty years ago, tracing the residence in 1881 of those born in a single parish in Dorset to establish the places elsewhere in England that had connections with that parish took several days work in an archive, and even then could only be achieved in part: see A. Hinde and M. Edgar, ' "Following the tools": migration networks among the stone workers of Purbeck in the nineteenth century, in M. Hammond and B. Sloan (eds), *Rural–Urban Relationships in the Nineteenth Century: Uneasy Neighbours* (London, 2016), pp. 90–104.

[24] Marshall, *Tyranny of the Discrete*, p. 97.

[25] Marshall, *Tyranny of the Discrete*, p. 99.

[26] Marshall, *Tyranny of the Discrete*, p. 98.

[27] Marshall, *Tyranny of the Discrete*, p. 100; Dennis Mills's comments were in Local Population Studies Society *Newsletter* 10.

the particular features of a place influence aspects of its demography. To shed some light on this is possible, and useful, even though we only have a nebulous idea of what the essence of the concept of 'place' is. Local studies have long established strengths in the identification of the mechanisms through which the local context affects population change. In the next section I show how the analysis of 'big data' can help identify *the extent* to which the unique features of particular places are important.

The place of 'place'

Demographic behaviour in the past varied across geographical space. Infant mortality was generally higher in towns than in the countryside; the mean age at marriage in the later nineteenth century was higher in the west of Wales than in most of England; fertility was higher in coal mining districts than in towns on the south coast of England.[28] What might be the role of 'place' in explanations of these variations? We can illustrate the arguments using an example, that of fertility in England and Wales between 1851 and 1911 taken from a recently constructed 'big data' set. The data set consists of local measures of fertility and marriage patterns, and is based on an analysis of individual-level census data made available through the I-CeM project mentioned earlier.[29]

The Cambridge Group for the History of Population and Social Structure's PopulationsPast web site displays the data in the form of maps of marital fertility, the average number of children per woman, the mean age at marriage and the proportion of women never married at ages 45–54 years for each of the census years from 1851 to 1911 (except for 1871) for more than 2,000 local geographical units in England and Wales.[30] If we take the year 1881 as an example, the average number of children per women (the total fertility rate in PopulationsPast terminology) ranged from more than 6.0 in parts of county Durham, south Yorkshire, Wigan in Lancashire and areas of the west Midlands to less than 4.0 in parts of western Wales, the Scottish border districts, London, middle class towns and some districts in Lancashire.[31]

Demographers know that in populations where fertility outside marriage is rare (which was true of England and Wales in 1881), overall fertility is largely determined by the proportions of women in the reproductive age groups that are married and the fertility of

[28] These examples are all illustrated on the PopulationsPast web site. See Cambridge Group for the History of Population and Social Structure, 'PopulationsPast: an interactive atlas of Victorian and Edwardian population', *Local Population Studies*, 100 (2018), pp. 77–81, https://www.populationspast.org/about/ [accessed 4 June 2019.

[29] E. Garrett and A. Reid, 'Composing a national picture from local scenes: new and future insights into the fertility transition', *Local Population Studies*, 100 (2018), pp. 60–76.

[30] A.M. Reid, S.J. Arulanantham, J.D. Day, E.M. Garrett, H. Jaadla, and M. Lucas-Smith, *PopulationsPast: Atlas of Victorian and Edwardian Population* (Cambridge, 2018), https://www.populationspast.org/ [accessed 5 June 2019].

[31] The highest and lowest values were 6.45 in Easington, County Durham, and 2.95 in Christchurch on the south coast. See Reid *et al.*, *PopulationsPast*. https://www.populationspast.org/tfr/1881/#6/54.072/-2.867 [accessed 5 June 2019].

those who are married.[32] Looking at the maps of these variables, it becomes clear that fertility within marriage varied rather little across England and Wales.[33] In the vast majority of the country it ranged between seven and nine children per woman.[34] Certain Lancashire textile districts, Sheffield, and a few south coast districts had lower values between 6.0 and 7.0. Only in central London were values below 6.0 observed. A cluster of four registration districts in the west of Wales had higher values over 9.0.

The mean age at marriage for women varied from below 25 years in the coalfields, heavy industrial areas and certain rural areas in eastern England and Kent to more than 28 years in the west of Wales and a couple of rural districts in the north of England.[35]

Those, then, are some of the patterns. The next question is how to explain them. Three or four decades ago, a common account would have suggested that geography, *per se*, was of little importance. Fertility and marriage patterns varied principally by occupation and social class, and geographical variations were manifestations of the fact that different areas had different social and occupational compositions. So, for example, fertility was low in prosperous towns like Harrogate and Brighton mainly because middle class people had low fertility and Harrogate and Brighton had a high proportion of middle class inhabitants. Similarly, fertility was high in county Durham because coal miners married young and had high fertility and the population of county Durham consisted of a high proportion of coal miners. To be sure, the social and occupational structure would not *completely* explain the fertility of a given area, but it was a major part of the explanation.

Factors such as the social and occupational structure, the argument runs, are measurable characteristics of every place, and have a general association with fertility. They are *ubiquitous* features. Many such features exist. If we were trying to account for geographical differentials in mortality in Victorian England and Wales, population density or the proportions of the population living in urban areas might constitute such features.[36]

The argument suggests, then, that a demographic phenomenon—such as the fertility we observe—in a given area is the sum of three components: the average fertility in the population as a whole, an adjustment to take account of ubiquitous features such as the social and occupational structure, and a 'residual' effect. This is true whether we are considering fertility in the aggregate or at the individual level. The residual encompasses all

[32] For the illegitimacy ratio, see Reid *et al.*, *PopulationsPast*, https://www.populationspast.org/ileg_ratio/1881/ #6/54.072/-2.867 [accessed 5 June 2019]. Fewer than 10 per cent of births in England and Wales in 1881 took place outside marriage.

[33] Reid *et al.*, *PopulationsPast*, https://www.populationspast.org/tmfr/1881/#7/53.035/-2.895 [accessed 5 June 2019].

[34] Fertility within marriage is measured using the total marital fertility rate, which is calculated for each registration district. This is (for most women) a hypothetical measure. It is the number of children a woman would have if she married at exact age 20 years, remained married until her 50th birthday, and at each age between 20 and 50 years had children at the rate pertaining to married women of that age in that district.

[35] Reid *et al.*, *PopulationsPast*, https://www.populationspast.org/f_smam/1881/#5/54.008/-8.240.

[36] See R. Woods, *The Demography of Victorian England and Wales* (Cambridge, 2000); and A. Hinde and B. Harris, 'Mortality decline by cause in urban and rural England and Wales, 1851–1910', *The History of the Family*, 24 (2019), pp. 377–403, https://doi.org/10.1080/1081602X.2019.1598463.

the other factors that we have not included in our adjustment for ubiquitous features that affect fertility. There are two groups of these. First there are general ubiquitous features that we could (or should) measure but have not done so in this case (examples might be income, or female autonomy). Second, there are 'place-specific' or 'individual-specific' features unique to each locality or each individual, the effects which are not captured by the general factors we have measured.

Not all 'place-specific' effects are in this residual. The impact of some may already be subsumed within the occupational structure variables. So, a feature specific to county Durham that is relevant for explaining its fertility levels in 1881 is that the land beneath it contained coal seams which were suitable for mining. Once that has been established, general (i.e. non place-specific) explanations of the social structure of mining populations take over and account for the early female marriage and the high fertility.[37]

Other 'place-specific' effects, however, may be less easy to accommodate within general accounts of the determinants of fertility, and might demand explanations of fertility which are unique to particular places, and which are properly the purview of local population studies. A key question, therefore, is how large and pervasive these kinds of 'place-specific' effects are.

At one extreme we might imagine that the observed levels of fertility in each place were entirely a consequence of features unique to that place. Thus the explanation of the fertility level in a place lies entirely within that place. In such a world all population studies would be local population studies. There would be no point in doing any analysis at a level beyond the individual locality, as such analysis would yield no new knowledge. At the other extreme we can imagine a world governed entirely by general associations between ubiquitous variables and fertility. If we could measure the impact of all these ubiquitous variables on fertility, then by knowing the values of these variables for each place we would know the fertility rate. In such a world there would be no place for local population studies.

Some of the geographical patterns we observe in fertility in England and Wales in 1881 are clearly explained in large part by the social and occupational structure: the high fertility in county Durham and the low fertility in middle-class towns. But other patterns are not. For example, in rural areas in the west of Wales, low fertility is mainly explained by late marriage and celibacy, even though women who do marry have high fertility within their marriages. This is an unusual combination of features which appears only in this area. In other rural areas for example in East Anglia quite different patterns are observed.

Of course, we could try to examine other ubiquitous social and economic variables across the whole data set to see how much of the geographical heterogeneity we could account for. This might include systematic variations across space in the way that some

[37] For some of these explanations, see D. Friedlander, 'Demographic patterns and socio-economic structure of the coal-mining population of England and Wales in the nineteenth century', *Economic Development and Cultural Change*, 22 (1973), pp. 39–51; and M.R. Haines, *Fertility and Occupation: Population Patterns in Industrialization* (New York, 1979). Similarly, fertility was low in predominantly middle class towns for reasons general to the middle classes: see J.A. Banks, *Prosperity and Parenthood: a Study of Family Planning among the Victorian Middle Classes* (London, 1954).

social and economic variables were associated with fertility. In theory, there may be almost no limit to the number of different factors we can add to an analysis of the determinants of fertility using the whole data set, and this will be true whether we use aggregate-level or individual-level data. In the end, it may be that we could thereby produce an explanation of the vast majority of local and regional patterns in terms of ubiquitous features, leaving no or only a very small residual effect.

'Big data' can act as a temptation to try to construct and measure the parameters of such general accounts of population change. Because 'big data' produces enormous numbers of cases, we can carry out exceedingly complex analyses to try to obtain results which are statistically significant, and to smooth over, or explain away, all the local and regional variations in whatever outcome we are analysing. We might try this, but there are several arguments against it. Taken together, these arguments reveal that, far from crushing local population studies beneath its weight, 'big data' has the potential to render local population studies more important and incisive than it has been hitherto.

'Big data' and the case for local population studies

Suppose we do try to develop a complicated general explanation of a demographic phenomenon such as fertility in England and Wales in 1881 using an analysis based on the whole population. The first problem is that the 'big data' set may not include all the variables we require.[38] 'Big data' sets in historical demography are typically constructed from census and vital registration data, and hence necessarily include a limited number of variables.

Even if we did have a huge range of potential ubiquitous covariates and could construct and quantify an imaginative and comprehensive explanation without reference to local factors, we might ask whether this is the most satisfactory way to understand the phenomenon we are analysing. Perhaps there is an 'optimal' explanation which lies between the two extremes of treating every locality as having its own story, and regarding all social processes as being governed by universal associations. This is, of course, just the old debate between idiographic and nomothetic explanations of social phenomena. However, recent developments in statistical learning designed for analysing 'big data' sets suggest not only that there is such an optimum, but provide potential statistical methods to identify it.[39] One key element in all this is applicability of the conclusions outside the data set which was used

[38] Even if we allow ourselves to create new ways of looking at the existing variables, perhaps by categorising occupation in different ways, or adding new proxy variables, such as those measuring latitude, longitude or climate.

[39] Jones *et al.*, *Introduction to Statistical Learning*, provide a good summary of these, with many examples. They demonstrate clearly that, if your aim is to construct an algorithm to predict the value of some quantity for a population *other than that used to estimate the parameters of the algorithm*, increasing the complexity of the algorithm does not always improve the accuracy of the predictions and, beyond a certain point, can cause it to deteriorate. In the statistical learning literature such a situation is called *over-fitting*. Another way of putting this is to say that increasing the number of variables in the analysis will lead to an increasing likelihood that some of associations identified will be false, especially if a range of statistical tests are used. See, for example, D. Spiegelhalter, *The Art of Statistics: Learning from Data* (Harmondsworth, 2019).

in the analysis leading to those conclusions. For example, suppose we developed an account of the geography of fertility in England and Wales in 1881 which explained everything in terms of associations between social class, occupation and fertility, perhaps with some observations which 'explained' variations with rural areas (for example sheep farming being associated with later marriage). This account was complicated, but was couched in general terms, leaving very little room for particular local effects. Would the same account provide a good explanation of the geography of fertility in 1891, or 1901, or 1911? The statistical learning evidence is quite clear that, beyond a certain level of complexity, it probably will not.[40]

One way to identify the variables whose associations with fertility can truly be regarded as general and wide-ranging might be to include in this category only those variables that have a strong association with fertility across many (or all) the census years for which an analysis was done. Another might be to compare the results of population-level studies from 'big data' sets in different countries, such as those available through IPUMS.

The point is that it is highly likely that a reasonable analysis of a demographic phenomenon in terms of general associations, no matter how 'big' the data set on which it is based, will probably leave quite a lot of geographical variation unaccounted for. Moreover, these 'big data' sets will allow the researcher to see exactly where (that is, in which localities) the unexplained variation is greatest. Here, then, are the strongest 'place' effects. Here are the localities where local studies have most to offer towards explanations of demographic phenomena. So, in the case of fertility in England and Wales in 1881, an area which stands out from the rest is the west of Wales, mainly in Cardiganshire (the present-day Ceredigion). In this area, as we have pointed out, fertility within marriage remained remarkably high until early in the twentieth century. There were other unusual features about this area, too. The ratio of males to females in the working ages was very low, with only about 60 men per 100 women in registration districts such as Aberayron in 1881.[41] In additional to having its own particular fertility regime, this was an area of high male out-migration, but how these features were associated is not clear, and is best understood through intensive local study. As pointed out by Tony Wrigley in his contribution to the hundredth issue of *Local Population Studies*, in social, economic and demographic history 'much of critical importance to appreciating the nature of … change … is concealed if the discussion focuses on exclusively on the whole country'.[42]

We have reached this point in the discussion saying little or nothing about the 'essence of place'. It may be that the local studies will uncover something of the latter. But even if they do not, we can conclude at a minimum that some unique features of some localities are important for understanding population structure and change, and that these are best researched at the local level.

[40] There are statistical tools to identify the optimum level of complexity for such a general account. See Jones *et al.*, *Introduction to Statistical Learning*, pp. 203–59.

[41] Reid *et al.*, *PopulationsPast*. https://www.populationspast.org/sr/1881/#12/51.9853/-4.7191 [accessed 11 June 2019].

[42] E.A. Wrigley, 'The general and the particular', *Local Population Studies*, 100 (2018), pp. 25–32, here at p. 25.

Other examples of local peculiarities have been revealed by ‘big data’. Kevin Schürer and Joe Day, in their recent analysis of migration patterns to and from London in the late nineteenth century, identify a general north-south divide along a line from the Wash to the Severn, to the north and west of which communication with London was less common than it was to the south and east. Places further away from London interacted less with the capital than did places closer to London, but, as they write: ‘the general model of distance decay is not totally borne out. In all the census years a number of seemingly isolated places some distance from the capital sometimes the same places over time in central-west Wales, in Cumbria and Northumbria, but also elsewhere, record a high proportion of natives who found their way to London.’[43] What was it about the inhabitants of these places that attracted them to London? Again, local population studies would seem the way to find out.

Symbiosis between ‘big data’ and local population studies

‘Big data’ therefore, has the potential to help local historians navigate one of the dangers of local study: the tendency to seek out and to write about the uniqueness of each place, even when places are not that unusual. Used as suggested in the previous section, ‘big data’ can shine a light on all localities, allowing the research to see which places really are different from the rest, or characterised by unusual demographic features. ‘Big data’ can help local historians and local demographers direct their attention to those places where local studies have their greatest contribution to make to knowledge.

Of course, to some researchers interested in pursuing local population studies, this will not seem a great advantage. These researchers are interested in their own local ‘place’, and that place may prove not to have any obvious outstanding features. But this need not render the local study of that place valueless. Because ‘big data’ is selective about the information it supplies, it is unlikely to be able to substitute for the deep contextual knowledge of local studies. In this sense, it has changed rather little, save to make the collection and processing of basic data easier.[44]

Neither will ‘big data’ affect the importance of studying localities where unusually detailed data survive. For some places, the existence of unique sources, or the availability of sources which are have not generally survived to the present day, allows the historian to paint a much richer picture of demographic processes than is the case in most places. Consider, for example, the operation of the New Poor Law in nineteenth century England. Some records are widely available (for example workhouse populations on the night of each census, or the numbers of paupers being relieved on 1 January and 1 July in each

[43] Schürer and Day, ‘Migration to London’, pp. 50–1.

[44] Another caveat with many large, publicly available, data sets is that access is granted not to the raw data, but to data that has been processed by the teams assembling the data. It is important to read the documentation accompanying the data sets to understand what that processing has involved. For some applications, historians may have to go back to the raw data even where processed versions of the same data are freely available.

year).[45] These, however, are but snapshots of the situation in each place at certain points in time. To ask questions about the extent to which the New Poor Law was used to manage the local labour market, its impact on migration, or the extent to which the poor could influence the operation of the poor relief system, recourse must be had to less widely available records, such as outdoor relief registers or workhouse admission and discharge registers.[46] The survival of these is very patchy, and these will constrain the localities in which the analysis is done, for only for these localities can the interesting questions be answered. What 'big data' can do is to save the historian time in obtaining basic data for these localities.

So local historians are unlikely to be put out of business by academic historians who now have access to increasing amounts of individual-level data for every place in the country. 'Big data' will make their lives easier, and to that extent they would do well to make use of the increasing amount of publicly available data sets. But it will not substitute for local and regional knowledge, nor for detailed examination of local sources. By highlighting unusual and particular features, 'big data' will generate new and interesting questions about the populations of certain regions and localities, which local population historians will be well placed to answer.

Acknowledgements

I should like to thank the other members of the *Local Population Studies* Editorial Board, and especially Heather Falvey and Chris Galley, for their comments on a draft of this paper.

[45] The numbers of paupers relieved on 1 January and 1 July in each year are given in the British Parliamentary Papers. See, as an example, Poor rates and pauperism, *Return (B.) – Paupers Relieved on 1st January 1862*. British Parliamentary Papers 1861 LIII [C. 324].

[46] The Old Poor Law had been routinely used to manage local labour markets. For an example of a local study demonstrating this, see B.K. Song, 'Landed interest, local government and the labour market in England, 1750–1850', *Economic History Review*, 51 (1998), pp. 465–88, https://doi.org/10.1111/1468-0289.00102.