Names | A Journal of Onomastics

A Note on the UK Local BMD: A Full Name Onomastic Resource

Stephen J. Bush
Xi'an Jiaotong University, Shaanxi, CHINA

ans-names.pitt.edu
ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 2, Summer 2024 DOI 10.5195/names.2024.2543

Abstract

Data from the UK Local BMD, a volunteer project to transcribe the birth, marriage and death records of England and Wales, is a rare onomastic resource, being one of the few public datasets to contain full names. However, it has yet to be presented in a form amenable to large-scale analysis. This article processes 25,213,860 birth and 9,887,244 death records—collectively representing 204,427 names across 289 years–into a resource for community use. The data are presented alongside a number of summary statistics and both internal and external validation of its integrity. The data, along with the code used to generate it, are available at http://www.github.com/sjbush/uk_bmd for non-commercial research purposes.

Keywords: first name, full name, England, Wales, BMD

Introduction

The UK Local BMD project (www.ukbmd.org.uk/local) is an ongoing volunteer-run effort to transcribe the birth, marriage, and death (BMD) records of England and Wales for digital preservation.[1] From an onomastic perspective, this is a rare and valuable dataset because of its temporal duration (the records begin in 1837[2]) and because it records full names. By contrast, many onomastic datasets, including those released annually by the UK Office for National Statistics, are limited only to the first name and have rarer names redacted. While the history of the local BMD data has been previously discussed (Bush, Powell-Smith, and Freeman 2018) and aspects of it used to explore specific onomastic phenomena (Bush 2019; 2020), the dataset as a whole has yet to be presented as a resource for the wider community.

To address this issue, the full set of birth and death records—as of September 2023—were processed into an extensive set of tables containing a number of summary statistics per name.[3] This article introduces the content of these tables, their internal and external validation, and the limitations of the dataset. It is hoped this data will not only complement existing English onomastic datasets but facilitate further research into the evolution of name use over time.

Results

Structure of the Raw BMD Dataset

The UK Local BMD is a composite of regional projects: at the time the birth and death records were downloaded (www.ukbmd.org.uk/local accessed September 13, 2023), it comprised the 12 cities, counties, and regions of Bath, Berkshire, Cheshire, Cumbria, Kingston-upon-Thames, Lancashire, North Wales, Shropshire, Staffordshire, West Midlands, Wiltshire, and Yorkshire. Their respective websites only permit the bulk download of 25 years' worth of records at a time for each letter, that is, a subset of records with surnames beginning with 'A', and so forth. Accordingly, 4,003 files had to be individually downloaded, representing each letter, year, and region (representing 2,001 files of birth and 2,002 files of death records, respectively). This raw dataset is the basis of subsequent processing, as detailed below.

The available fields for each birth record were the forename(s) and surname, mother's pre-marital surname (where applicable), year of birth, sub-district of the region in which the birth was registered, and reference number. Records from Shropshire and Wiltshire did not contain complete middle names but only initials. The available fields for each death record were the forename(s) and surname, age at death (where known), year of death, sub-district of the region in which the death was registered, and reference number. As with the birth records, death records from Shropshire and Wiltshire only contained middle-name initials. Records from Cheshire and Staffordshire were also an exception, as these listed either age at death, or—in the same field—the date of birth, if known. Using the year of death and age at death, we imputed the year of birth and thereby created two comparable sets of yearly name use data, 'B' and 'D', which we used to assess the completeness of each other (discussed below). Unfortunately, age at death was not provided for records in Cumbria, Shropshire, and the West Midlands, and as such, no death records could be used from those regions.

Spaces in the forename field were assumed to separate individual names. The name before the first space was considered the first name and all subsequent names, space delimited, were middle names. For example,

DOI 10.5195/names.2024.2543 Vol. 72 No. 2, Summer 2024 ISSN: 0027-7738 (print) 1756-2279 (web)

Ellen Sarah Jane has the first name *Ellen* and two middle names, *Sarah* and *Jane*, but *Sarahjane Ellen* has one first name, *Sarahjane*, and one middle name, *Ellen*.

To avoid overcounting the same name, we also required that there be only one record per reference number. Multiple records with the same reference typically (but not always) denote variants of the same name, such as those without clear distinctions between the 'forenames' and 'surname' fields (for example, a pair of records with the names *Tyler Van Der KAMP* and *Tyler VANDERKAMP*, where the surname field is capitalised). However, this was only the case for Bath, Cumbria, Shropshire, the West Midlands, Wiltshire, and Yorkshire. In the other regions, reference numbers could not be used as they were not individual; rather, they referred to a processing batch of generally 5–10 people at a time. (A table giving the number of usable records per region and the years and regions covered is available at http://www.github.com/sjbush/uk_bmd). All subsequent references to data, code, and summary statistics refer to files at this location.

Limitations of the Dataset

There are unavoidable limitations to using a volunteer-collected dataset. It is in the nature of volunteering that efforts are discontinuous, with several of the regional projects not being actively maintained, having no new birth or death records for over a decade (detailed in the file 'summary_of_records_per_region.txt', described below). Nevertheless, as records have been transcribed on an *ad hoc* basis, we may still assume it is a representative—if incomplete—population sample. This is assessed more formally below.

Previous uses of this data (Bush, Powell-Smith, and Freeman 2018) have noted the presence of typographical errors, unexpanded abbreviations (such as *Wm* for *William*), non-names included in the name field (such as births registered as *Un-named* or deaths as *Unchristened*), descriptions in the name field (such as the 'forenames' *John Mother's Name Jane*), and names in the middle position which end in a hyphen (which are likely the first part of a compound surname). Although the forename field was parsed to exclude or edit a small number of records of this kind (criteria detailed at www.github.com/sjbush/uk_bmd), we considered that the correction of typographical errors would not only be subjective but unlikely to be comprehensive. Accordingly, a number of typographical errors will inevitably remain in the final, processed, dataset. Nevertheless, we reasoned that by making the raw data available as is, with minimal editing, we avoid constraining what others may later do with the dataset. We also avoid making problematic assumptions as to what a given name "should" be, if it was "correct" (for example, by revising *Sarahjane* to *Sarah Jane*).[4] The forenames field was also parsed to exclude those records which did not contain at least one alphabetical character, those which contained numbers, and those which denoted ambiguity (for instance, by including a question mark or by placing the name in brackets to reflect a lack of confidence in the transcription).

Aside from regional heterogeneity, the most obvious limitation is that the BMD does not record the gender of the individual (more precisely, sex assigned at birth), hindering analysis of this fundamental property. Accordingly, we predicted the gender of each name using the US Social Security Administration dataset, as in West et al. (2013). This dataset, comprising 365,296,191 records from 1880 to 2022 (a similar range to the BMD), contains the first names and gender of all US Americans with a social security number, with names registered to fewer than 5 people a year excluded (https://tinyurl.com/2aumvhvk, accessed 11 June 2023).

We gender-typed a name as male or female only when it was more frequently assigned to that gender for every year in which it was recorded (noting that this dataset acknowledges only two genders). This is a pragmatic approach to predicting gender although is freighted with complications, being a "blunt tool to study a complex subject" (Blevins and Mullen 2015). Most notably, gender associations are not timeless. Accordingly, names which changed their primary gender over time (such as *Madison*, generally a male name before 1984, and female after) were sub-classified as 'mostly male' and 'mostly female' as appropriate, on the basis of the total number of records. Although the US SSA is a large dataset with broad temporal scope, there are nevertheless many names in the UK BMD that are not included in it. To account for this where possible, we repeated the above gender-typing strategy using three additional datasets, each of which were full population samples of all live births in (a) England and Wales from 1996 to 2021 (15,620,686 records), sourced from the UK Office for National Statistics (ONS) (https://tinyurl.com/58b5auyt, accessed 14 June 2023), (b) Scotland from 1974 to 2022 (2,918,610 records), sourced from the National Records of Scotland (https://tinyurl.com/324e2evb, accessed 14 June 2023), and (c) the province of Alberta, Canada, from 1980 to 2020 (1,840,539 records), sourced from the Government of Alberta (https://tinyurl.com/5n7vzy7x, accessed 14 June 2023).[5]

Like the US SSA dataset, but unlike the UK BMD, the UK ONS, NRS and Alberta datasets do not include middle names or surnames. Similarly, the UK ONS dataset excludes names registered to fewer than 3 people a year. The NRS and Alberta datasets, however, do not stipulate a minimum number of births per name.

By means of pooling the count data from all four datasets, gender classifications were made for 211,627 names of which 77,786 were male, 5268 'mostly male', 122,548 female, 4,799 'mostly female', and 1,226

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 2, Summer 2024 DOI 10.5195/names.2024.2543

'undetermined' (because there was an even number of male and female records). Names not present in this pooled dataset could not be automatically gender-typed; in this case, the default gender classification is 'undetermined'.

Structure of the Processed BMD Dataset

The raw dataset of 4,003 files were processed using custom scripts to produce the following tab-separated plain text files. For each of the 'B' and 'D' raw datasets, the processed data files were

1. summary_of_records_per_region.txt

   A table listing the regions covered, the URL and date of last update for the raw data, the total number of records processed, and the years covered

2. summary_of_records_per_year.txt

   A table listing, per year, the following summary statistics:

   a. The total number of records with a first and middle name

   b. The total number of records with male- and female-typed first names. Male names are those classified as 'male' or 'mostly male' and female as 'female' or 'mostly female', according to the above definition.

   c. The male/female gender ratio of the first name records (A proxy for how representative the data is, discussed further below.)

   d. The mean number of middle names per birth record (for birth records with ≥ 1 middle name)

   e. The number of different first names seen

   f. A measure of 'forename diversity' (the ratio of the number of different first names to the total number of records per year)

   g. The percentage of that year's first names not recorded in the previous year

   h. The top five most popular first and middle names

3. all_first_names.txt, all_middle_names.txt, and all_first_and_middle_names.txt

   A list of all first names, middle names, and first name-middle name combinations recorded, their associated gender, and both their absolute and relative (percentage) count

4. all_full_names/X.txt, where X is one of the 12 local BMD regions.

   A list of all full names recorded in that region, with their associated year of birth

5. first_name_as_percentage_of_records_per_year.txt,
first_name_as_absolute_number_of_records_per_year.txt,
middle_name_as_percentage_of_records_per_year.txt.txt, and
middle_name_as_absolute_number_of_records_per_year.txt

   Four tables giving, for each first and middle name, their absolute and relative (percentage) count per year. For each name, the following summary statistics are also given:

   a. The total number of records, across all years, with that name

   b. The percentage of the total number of records, across all years, with that name

   c. The gender of that name, considered either 'male', 'female', 'mostly male', 'mostly female' or 'undetermined'

   d. The total number of years, and maximum number of consecutive years, in which each name is registered

   e. The years of first, last, and peak recorded use of that name (as the maximum percentage of all births registered in a given year)

   f. The year in which the name is considered 'abandoned', if applicable. This is defined as the year in which the proportion of births with that name first drops below 10% of its past maximum, as in Berger and Le Mens (2009). We assess whether this is the case only for names used (1) a total of > 1000 times, and (2) in consecutive use for ≥ 50 years[6], with (3) their peak usage within the longest consecutive use period. This definition of 'abandonment' draws on the fact that the majority of names have one peak of popularity from which decline is consistent (unless usage of that name is later revived) (Berger and Le Mens 2009).

DOI 10.5195/names.2024.2543 Vol. 72 No. 2, Summer 2024 ISSN: 0027-7738 (print) 1756-2279 (web)

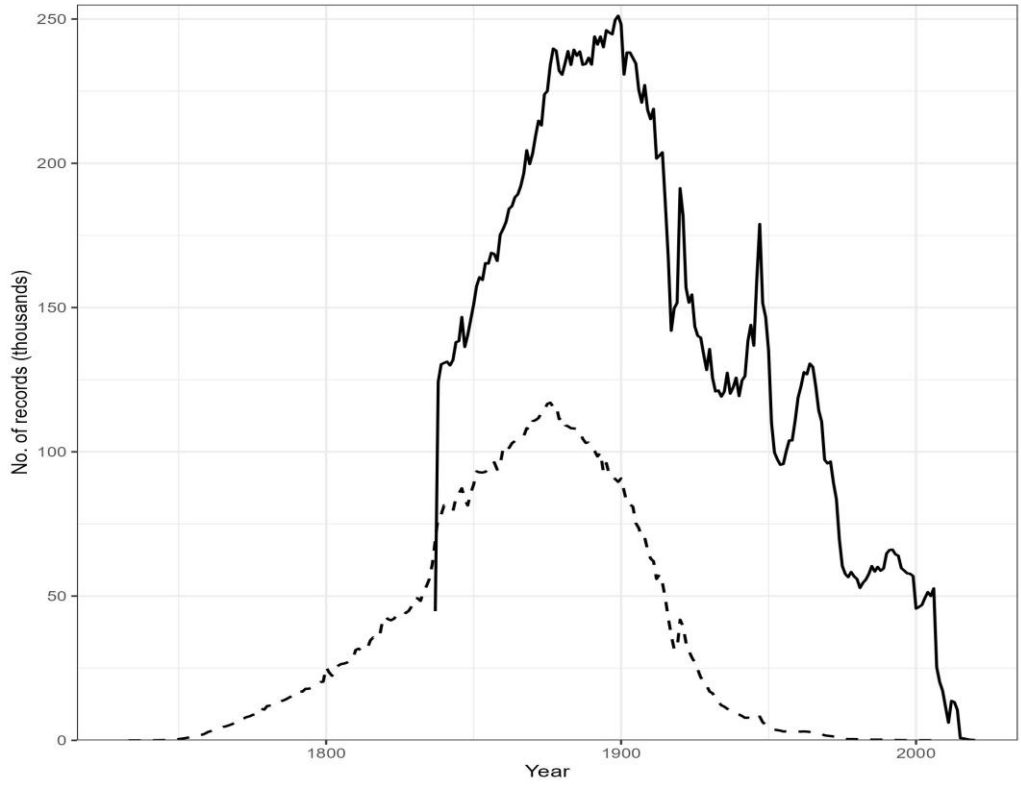
g. The year in which an abandoned name is considered revived. If after dropping below 10% of its past maximum, a name is revived if after 5 or more years, its use is > 25% of its past maximum. Names are only considered revived if the year of revival occurs within the longest consecutive use period, that is, the name cannot have vanished from the record completely (that being more likely to indicate sparse record-keeping).

h. The number of years between debut and peak use, debut and abandonment, peak and abandonment, and abandonment to revival (if applicable)

i. The rank order of that name, both overall (i.e. based on total number of records) and in the male and female categories.

Overall, the data comprised 25,213,860 birth and 9,887,244 death records (of which 44.66% and 26.98% had at least one middle name, respectively), spanning the years 1837–2022 and 1733–2009, respectively. Collectively, this represented 204,427 different names, of which 95,943 first and 70,535 middle names were only found in the 'B' dataset, and 11,326 first and 14,120 middle names only in the 'D' dataset, respectively. The number of records per year is shown in figure 1, and although variable, is generally in excess of 100,000 birth records per year for 130 consecutive years (the period 1838–1968). The number of death records per year, by contrast, peaks in the year 1878, with relatively few 20th century records (<10,000 per year from 1938 onwards) and negligible coverage at the extremes of the distribution (<100 records per year in the periods 1733–1744 and 1995–2009).

Figure 1: The Number of Birth (Solid Line) and Death (Dashed Line) Records Per Year



Internal and External Validation of the Data

To validate the data, we first compared the total number of records per forename in the 'B' and 'D' datasets (figure 2), finding a strong positive correlation between them (Pearson's $r = 0.976$, $p < 2.2x10^{-16}$ (Schober, Boer,

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 2, Summer 2024 DOI 10.5195/names.2024.2543

and Schwarte 2018)). This is a crude sanity-test of the data, and although not controlling for differences in either temporal or geographical scope, it nevertheless suggests that both datasets are, correctly, randomly sampling from the same population.

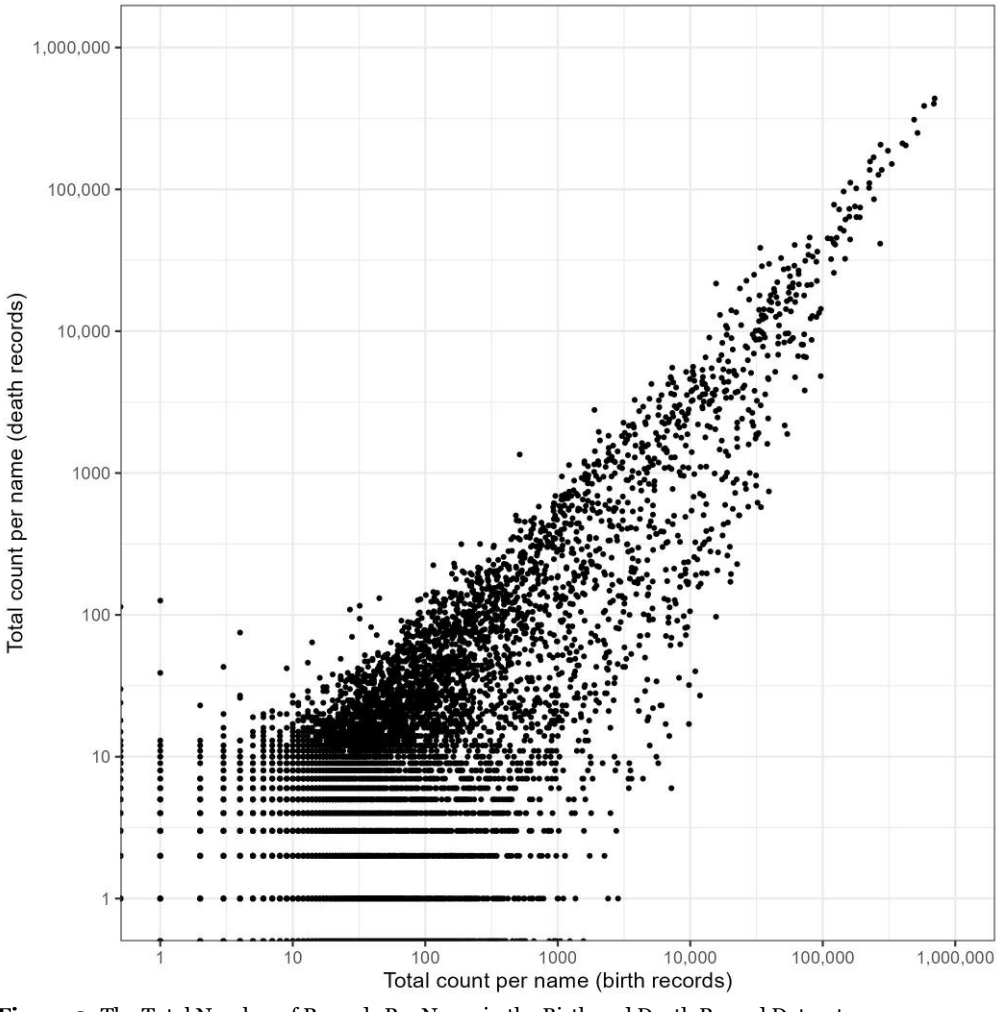


Figure 2: The Total Number of Records Per Name in the Birth and Death Record Datasets

A more nuanced validation was performed by comparing the total number of records per forename in the 'B' dataset to the total number of records per forename in the UK ONS dataset (figure 3). We restricted this analysis only to the years 1996–2007, as in this period the two datasets had a substantive number of records in common (the 'B' dataset has >10,000 records per year for each of these years). There was a very strong positive correlation between the 'B' and ONS datasets (Pearson's $r = 0.995$, $p < 2.2 \times 10^{-16}$), consistent with the fact that the former is in practice, and as expected, a subset of the latter. Unfortunately, it was not possible to compare the 'D' dataset with the ONS dataset as the former only contained 730 records within the period 1996–2009.

DOI 10.5195/names.2024.2543 Vol. 72 No. 2, Summer 2024 ISSN: 0027-7738 (print) 1756-2279 (web)

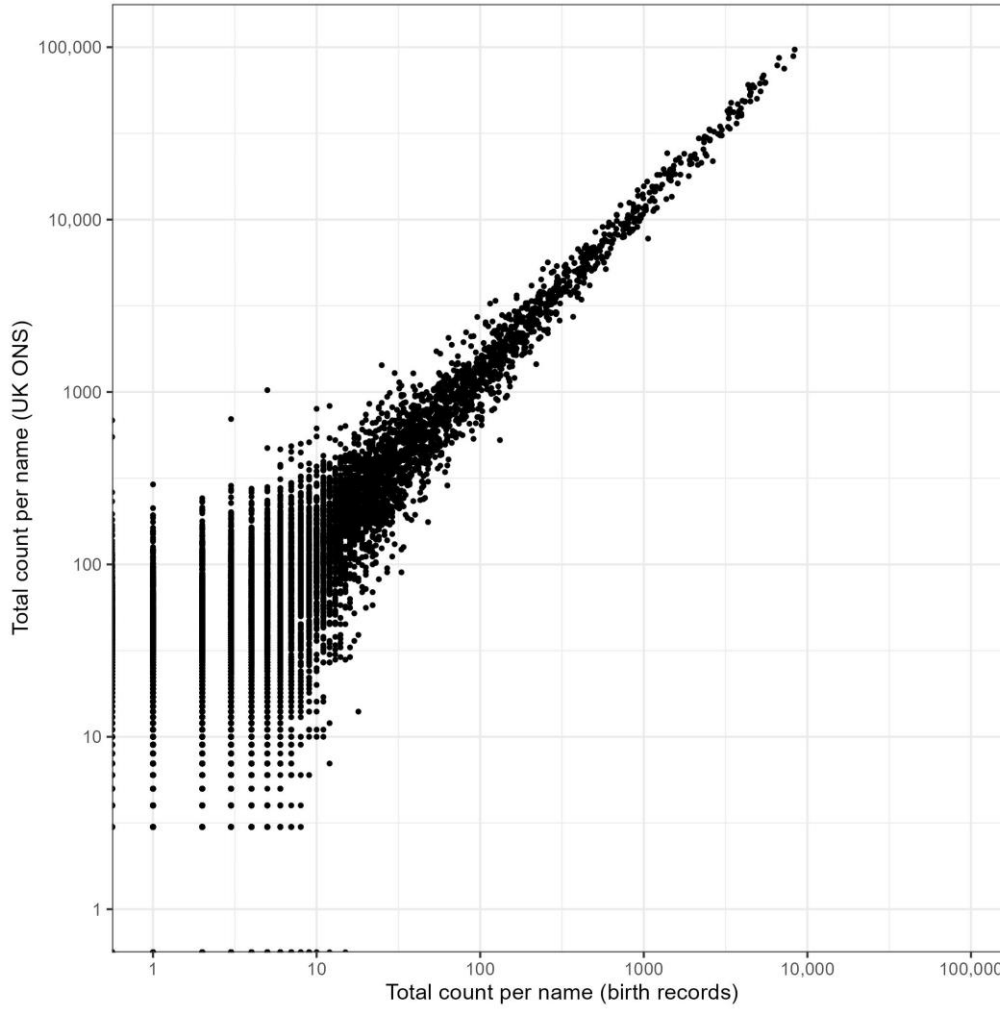


Figure 3: The Total Number of Records Per Name in the Birth and UK ONS Datasets, Restricted to the Period 1996–2007

Finally, we considered the gender balance of the data. The male/female (M/F) ratio of human births is consistently biased towards males and generally falls within the range of 1.03–1.06 males per female born (Chao et al. 2019). Deviations from this range can indicate possible biases in record-keeping or suggest that the data are not a representative population sample. Accordingly, we plotted the M/F ratio for the number of birth and death records per year (figure 4).

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 2, Summer 2024 DOI 10.5195/names.2024.2543

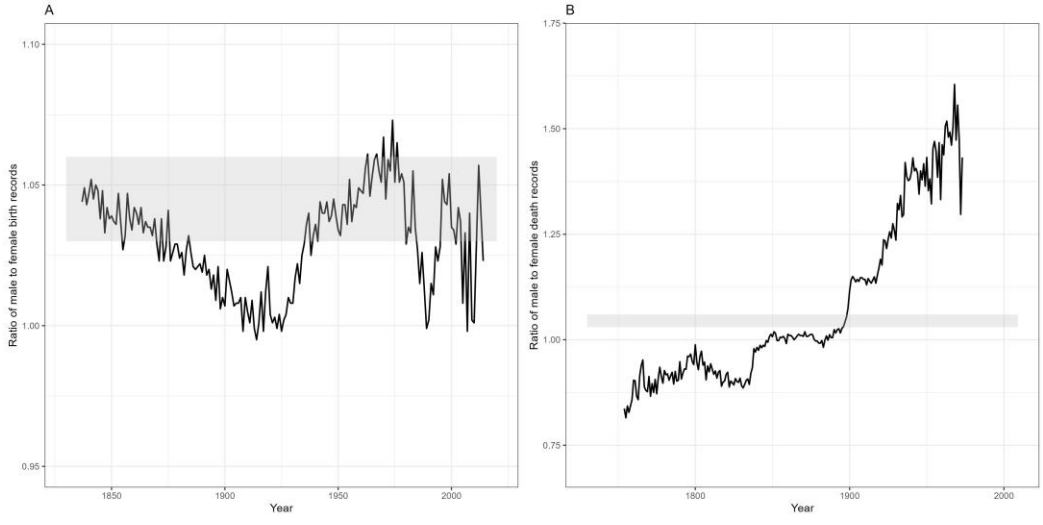


Figure 4: The Male/Female Ratio for (A) Birth, and (B) Death Records
Note: The shaded grey box denotes the expected birth ratio for humans: a range of 1.03–1.06 males per female born. Years with under 10,000 birth or 1,000 death records were excluded.

For the birth records, there was no deviation from the expected M/F ratio for the mid-19th and mid-20th centuries, suggesting the data are an unbiased sample. However, during the latter half of the 19th century and the first half of the 20th, the M/F ratio declines to parity before rising again. This is possibly because increasing rates of social change and immigration (among other factors) introduced many new names to the population–ones not included in the databases used to predict gender (Recall that the ONS, NRS, and Alberta datasets only begin in 1996, 1974, and 1980, respectively, much later than the 'dip' shown in figure 4A). As there is generally a larger pool of female than male names, this skewed M/F ratio may reflect the fact that female names (in the UK BMD) were less likely to be found in the gender-typing datasets.

Interpreting the gender balance of the death records is more complex, as there are a number of historic gender biases in mortality (for example, estimates of the death rate in childbirth in England range from 5 to 29 women per 1,000 by the late eighteenth century (Loudon 1986). Nevertheless, for those years with the largest number of records (approximately 1850–1900, as shown in figure 1), we can see that the male/female death ratio approached parity, suggesting that as with the birth records there was no systematic coverage bias for many years (figure 4B). The overall conclusion is that while the UK BMD datasets have a degree of heterogeneity, they remain generally representative of the population as a whole. These datasets are also rare resources by virtue of containing full names and so, it's hoped, they may further research into historic naming trends.

Disclosure Statement

The author declares that there are no conflicting interests.

DOI 10.5195/names.2024.2543 Vol. 72 No. 2, Summer 2024 ISSN: 0027-7738 (print) 1756-2279 (web)

Notes

[1] The name is somewhat misleading as the UK BMD contains no records from Scotland or Northern Ireland. This is due to differences in the legislative frameworks and history of civil registration relative to England and Wales.

[2] In both England and Wales, civil registration began on 1 July 1837, although it only became compulsory from 1 January 1875, with the passing of the Births and Deaths Registration Act 1874.

[3] The website hosting the UK Local BMD project (http://www.ukbmd.org.uk) is operated by Weston Technologies Limited (Crewe, Cheshire, UK). This company is the owner or license-holder of the intellectual property constituting the birth and death records, as detailed at https://www.ukbmd.org.uk/TermsAndConditions (accessed 7 June 2023). Under section 29A of the UK Copyright, Designs and Patents Act 1988, a copyright exception permits copies to be made of lawfully accessible material in order to conduct text and data mining for non-commercial research.

[4] *Sarahjane* is registered as a first name 11 times, so although rare, it is unlikely to be an error.

[5] Shortened URLs have been used here for readability. For the avoidance of error, the full URLs for each of the US SSA, UK ONS, Scottish and Canadian datasets are, respectively,

1) https://www.ssa.gov/OACT/babynames/names.zip

2) https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2021/relateddata

3) https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/names/babies-first-names

4) https://open.alberta.ca/opendata/frequency-and-ranking-of-baby-names-by-year-and-gender.

[6] These thresholds are arbitrary. They can be varied by re-running the scripts.

References

Berger, Jonah, and Gaël Le Mens. 2009. "How Adoption Speed Affects the Abandonment of Cultural Tastes". *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 20: 8146–50. https://doi.org/10.1073/pnas.0812647106

Blevins, Cameron, and Lincoln A. Mullen. 2015. "Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction". *Digital Humanities Quarterly* 9, no. 3. https://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html

Bush, Stephen J. 2019. "Re-Using the Names of Newborns: Symbolic Reincarnation in an Age of Infant Mortality". *Names* 67, no. 2: 100–112. https://doi.org/10.1080/00277738.2018.1536186

Bush, Stephen J. 2020. "Ambivalence, Avoidance, and Appeal: Alliterative Aspects of Anglo Anthroponyms". *Names* 68, no. 3: 141–55. https://doi.org/10.1080/00277738.2020.1775471

Bush, Stephen J., Anna Powell-Smith, and Tom C. Freeman. 2018. "Network Analysis of the Social and Demographic Influences on Name Choice within the UK (1838–2016)'. *PLOS ONE* 13, no. 10: e0205759. https://doi.org/10.1371/journal.pone.0205759

Chao, Fengqing, Patrick Gerland, Alex R. Cook, and Leontine Alkema. 2019. "Systematic Assessment of the Sex Ratio at Birth for All Countries and Estimation of National Imbalances and Regional Reference Levels". *Proceedings of the National Academy of Sciences of the United States of America* 116, no. 19: 9303–11. https://doi.org/10.1073/pnas.1812593116

Loudon, I. 1986. "Deaths in Childbed from the Eighteenth Century to 1935". *Medical History* 30, no. 1: 1–41. https://doi.org/10.1017/s0025727300045014

Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. "Correlation Coefficients: Appropriate Use and Interpretation". *Anesthesia and Analgesia* 126, no. 5: 1763–68. https://doi.org/10.1213/ANE.0000000000002864

West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. 2013. 'The Role of Gender in Scholarly Authorship'. *PLOS ONE* 8, no. 7: e66212. https://doi.org/10.1371/journal.pone.0066212

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 72 No. 2, Summer 2024 DOI 10.5195/names.2024.2543

Notes on Contributor:

Stephen J. Bush is a computational biologist who has had a personal and academic interest in names and naming for many years. His onomastic research applies the tools and techniques of bioinformatics to quantitative name data, centred at present largely on the UK BMD (birth, marriage, death) registers. As names are a product and reflection of cultural changes over time, he is ultimately interested in understanding how these came about, what factors influenced them, and how they spread.

Correspondence to: stephen.bush@xjtu.edu.cn

DOI 10.5195/names.2024.2543 Vol. 72 No. 2, Summer 2024 ISSN: 0027-7738 (print) 1756-2279 (web)