Journal of Retailing and Consumer Services 30 (2016) 271–278

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser

Deriving age and gender from forenames for consumer analytics

Guy Lansley, Paul Longley *

Department of Geography, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form
5 February 2016
Accepted 10 February 2016
Available online 3 March 2016

Keywords:
Age
Gender
Geodemographics
Big Data
Social media

ABSTRACT

This paper explores the age and gender distributions of the bearers of British forenames and identifies key trends in British naming conventions. Age and gender characteristics are known to greatly influence consumption behaviour, and so extracting and using names to indicate these characteristics from consumer datasets is of clear value to the retail and marketing industries. Data representing over 17 million individuals sourced from birth certificates and market data have been modelled to estimate the total age and gender distributions of 32,000 unique forenames in Britain. When aggregated into five year age bands for each gender, the data reveal distinctive age profiles for different names, which are largely a product of the rise and decline in popularity of different baby names over the past 90 years. The names database produced can be used to infer the expected age and gender structures of many consumer datasets, as well as to anticipate key characteristics of consumers at the level of the individual.

1. Introduction and overview

The advent of new sources of consumer data, such as those arising from the use of social media, online shopping and customer loyalty databases, present new opportunities to measure and model the activity patterns of individuals. Such data may be related to detailed functional taxonomies of the locations that consumers visit or the products that they buy, bringing insight into the nature and likely motivations for observed activity patterns. But there has been no commensurate improvement in the detail with which we are able to characterise the individuals themselves, and thus ascertain how representative they are of consumer segments or indeed the population at large.

In this context, our own attempts to understand consumer behaviour have become focused upon the task of front-loading the inferences that may be drawn from consumer names and social media user identifiers, in order to relate new Big Data sources to the wider populations from which they are drawn. In this paper we describe some of the ways in which individual given (fore-) names can be analysed in order to ascribe age and gender characteristics, as part of this task. The work builds upon a commercial classification (Monica: CACI, London) and is part of a wider research programme which explores given and family name pairings in order to infer individual characteristics in consumer research (Mateos et al., 2011; Longley and Adnan, 2016).

A person's given name can be used to infer a number of key individual characteristics, as a result of the ways in which names are typically distributed according to age, gender and ethnicity. Most forenames are gender specific and many can be traced to ethnic groups through common heritage (Mateos et al., 2011). There has been much research into patterns in names and how they link to cultural heritage and wider society. However, only limited attempts have been made to consider the age and gender distributions of forenames in Britain (Finch, 2008). Names vary by gender and age across time largely because of shifts in popularity of different baby names over time and the influence of migration. This paper summarises work undertaken to model the age and gender structures of different name holders in the Great Britain. From achieving a greater understanding about the demographic characteristics of different name bearers, more information can be inferred from consumer data which include names but no further details about the individuals.

Retailers have benefitted greatly from geodemographic datasets, made available from the government or other businesses, as a means of segmenting and understanding consumers. Such data allow retailers to plan their stock and marketing accordingly to the local population characteristics (Mitchell and McGoldrick, 1994; O'Malley et al., 1997). However, this conventional approach is based upon the assumed correspondence between night-time residence and consumer behaviour (Harris et al., 2005). The data is also typically aggregated and/or modelled, therefore the traits for each spatial unit may not be entirely representative of every resident (Openshaw, 1984). Consequently, there have also been developments in micro-economic modelling which seek to place the

* Corresponding author.
E-mail addresses: g.lansley@ucl.ac.uk (G. Lansley), p.longley@ucl.ac.uk (P. Longley).

http://dx.doi.org/10.1016/j.jretconser.2016.02.007

focus upon the individual to arguably provide an inherently superior approach to understanding consumer behaviour as it circumvents any issues of ecological fallacy (Hensher and Johnson, 1981; see also Longley and Adnan (2016)). However, in order to apply such techniques individual level data is required, and often consumer data is absent of demographic information. For instance, the records of an online account associated with large retailer will typically only include a name, address and purchase history. Therefore, inferring demographic traits from names data could allow analysts to harness more consumer insight from many data sources.

1.1. Demographics and consumption

In this paper we take it as axiomatic that gender and age of consumers both heavily influence behaviour and consumption practices. Consequently, much research has been devoted to understanding the influence of demographics on consumption, such as the incorporation of such characteristics in product and brand choice models (Kalyanam and Putler, 1997), and identifying target demographics for new products. Geodemographic segmentations are widely used to segment the population in to distinctive consumer groups from multivariate data (Harris et al., 2005), and age and gender are included in them as key correlates of consumer behaviour.

The gender divide in consumption practices is most obvious amongst certain product types such as clothes and cosmetics, where retailers produce and sell entirely different lines of stock tailored for each sex (Scanlon, 2000). Consumption has become an important means through which individuals construct their gendered identities (Baudrillard, 1998). There are even gendered variations in consumption of products which are not produced exclusively for one gender. For instance, women have been found to be more health conscious as exerted by their food consumption practices and therefore they are generally likely to perceive certain foods differently to men (Wardle et al., 2004). There is even a gender divide in perceptions of shopping behaviour, women being traditionally more likely to perceive shopping as a leisure experience and therefore spend more time visiting high streets (Campbell, 1997; Lunt and Livingstone, 1992). By contrast, males have been traditionally the most dominant patrons of online shopping websites (Dittmar et al., 2004; Rodgers and Harris, 2003), and are more comfortable using multiple channels when making a purchase (Blázquez, 2014).

Age is also an influential characteristic of consumer behaviour, in terms of product and brand preference, and also in terms of how individuals shop. For instance, younger consumers are usually more likely to patronise online shopping channels, whilst older individuals are typically less engaged with the Internet and other modern shopping channels such as mobile commerce (Sorce et al., 2005). Consumer behaviour is greatly influenced by the family life cycle and the ways in which disposable incomes are channelled through consumption (Reynolds and Wells, 1977). Amongst the adult population, different cohorts are known to have different consumption practices which link to their life stage, their physical and health characteristics, and their cultural characteristics. Shared experiences during adolescence and beyond that are traceable to societal, cultural and environmental traits can encourage individuals to develop values that they will retain over time and can give rise to “cohort effects” (Harmon et al., 1999). Consequently, those from an age cohort may share distinctive values and this is likely to influence their consumer behaviour (Pentecost and Lynda, 2010).

Understanding the demographic characteristics of consumers is therefore very important to developing a sustainable retail strategy. Consequently, inferring demographic traits of consumers has been a vital area of marketing and consumer research (McDonald, 1995; Carpenter and Moore, 2006). It is also important to adapt to local and national demographic changes. For instance, there is an increasing imperative to understand the consumption practices of elderly consumers in many western countries given their aging populations (Kohijoki and Marjanen, 2012). Even amongst stable populations, previous research has established that age cohorts have unique consumption traits relative to previous generations (Bakewell and Mitchell, 2003). Although every individual may have distinctive tastes, general consumption practices nevertheless vary by age and gender. Therefore, the possibility of estimating the general demographic structure alongside a consumer's distinctive personal characteristics from individual customer records can be a fruitful means of obtaining key information about clients and customers.

2. Names and demographics

There is a wide range of Big Data sources on the population which includes name identifiers, but have little or no additional demographic information. These include electoral registers, customer records and social media data. There have subsequently been attempts to harness information from names by examining how names are distributed through contemporary society. Perhaps the most sophisticated developments have been in the production of cultural, ethnic and language group classifications from forename–surname pairs (Mateos, 2007; Mateos et al., 2011). Surnames, in particular, can identify bonds between family members, and therefore can be aggregated to represent distinctive cultural groups.

Historically, there has been a range of processes that influence popular naming conventions in the Great Britain (Smith-Bannister, 1997). Over the last century, popular naming practices have become far more erratic (Galbi, 2002), reflecting secularisation of society, migration trends and social mobility. The UK Office for National Statistics (ONS) has nevertheless identified clear trends in baby naming over the years (Matheson and Summerfield, 2000), and names therefore offer a viable means of estimating age structures from larger populations (Scharf, 2005).

We associate names with their bearers, yet forenames more directly manifest the predilections, priorities and preferences of either or both parents (Gureckis and Goldstone, 2009). The choice of baby name is likely to vary systematically between parents from different socio-economic backgrounds and cultural groups. The favourability of names is also influenced by popular trends, giving rise to temporal autocorrelation in name frequencies (Xi et al., 2014). Research has identified that parents in the USA perceive baby names which are growing in popularity to be more desirable than those whose popularity is waning (Berger and Mens, 2009; Gureckis and Goldstone, 2009). In additional to shifts in societal values, more subtle environmental and internal influences also drive the popularity of names for particular groups (Lieberson, 2000). Whilst the choice of baby name is influenced by various sociological factors, some names are handed down by family members or have remained popular because of links to cultural heritage: such names are much less likely to vary much between age groups (Finch, 2008).

Forenames are subsequently an important part of an individual's identity and can even act as a positive source for cultural capital (Lord, 2002). Observers may associate names with stereotypes, such as a child's likely educational attainment (Harari and McDavid, 1973; Erwin and Calev, 1984). This may be grounded in truth because of the different forename preferences of parents who themselves have different experience of, and attitudes to, educational capital formation.

A major source of names data for Great Britain is the 44.8 million individual records that make up the 2011 CACI (London, UK) Consumer Register. This is built from a number of sources, including the public version of the national Register of Electors. Our analysis of this data source investigates how forenames are distinctively distributed between the super-groups that make up the 2011 UK Output Area Classification (OAC: Gale, 2014). Similar names associations have been developed in the past for commercial segmentations. The CACI Consumer Register used in this analysis was compiled for the same year as the most recent UK Census, which underpins the OAC classification. The database includes the forename and surname of registered individuals, and also recorded their home postcode so their residential locations could be identified. Forenames from the Consumer Register were joined to residential Census statistics at small area level. The elemental Census Output Areas that make up the OAC classification had an average population of just 309 for England and Wales in the 2011 Census, and are the smallest unit census estimates that have been publicly published under (ONS, 2015). The OAC segments output areas into 8 super-groups by a range of population statistics pertaining to demographics, cultural identity, socio-economics, employment and household characteristics. Table 1 shows some of the results obtained by aggregating the forenames from the Consumer Register to Output Area scale in order to link with the OAC. The table displays the top five most overrepresented names from each OAC super-group relative to the national average.

The two most ethnically diverse super-groups (namely Ethnicity Central and Multicultural Metropolitans) have a much higher representation of foreign origin first names. The Hard-Pressed Living group contain modern variants of some traditional names. Names popular in Victorian times are more common in the Suburbanites super-group, suggesting there is a link between socio-economic status and naming practices. It is also notable that the two super-groups with high proportions of elderly residents, Rural Residents and Constrained City Dwellers, are over represented by forenames that were more fashionable in bygone times. The Cosmopolitans super-group, which has a higher proportion of young adults, is characterised by shortened names.

Taken together, this analysis suggests two key findings: the popularity of different forenames varies between neighbourhoods; and many of these differences can be accounted for by local neighbourhood characteristics, notably cultural heritage and age. Whilst research into heritage and naming connotations is well established (e.g. Mateos et al., 2011), the influence of temporal popularity of forenames and the subsequent age structure of forename bearers in England and Wales is relatively under-researched. In the following section, we therefore generalise the age and gender structure of the bearers of given names in the UK, using data from birth certificate records to supplement the CACI Consumer register.

Table 1
The top five over-represented forenames for each 2011 OAC super-group, based upon.names with a frequency of 10,000 or more in the 2011 Consumer Register (Source: CACI, London).

Rural residents	Cosmopolitans	Ethnicity central	Multicultural metropolitans
PENELOPE	TOM	MOHAMED	MOHAMMED
HUGH	NICK	AHMED	MUHAMMAD
ALASTAIR	HARRIET	ALI	MOHAMMAD
ROSEMARY	MAX	JOSE	ABDUL
PHILIPPA	ALEX	ABDUL	AHMED
Urbanites	**Suburbanites**	**Constrained City Dwellers**	**Hard-pressed living**
TOBY	HILARY	LILLIAN	KAYLEIGH
PHILIPPA	GEOFFREY	MAY	LEANNE
JEREMY	KATHRYN	ETHEL	LYNDSEY
KATHERINE	JILL	KAYLEIGH	STACEY
DUNCAN	GILLIAN	ELSIE	KYLE

3. Enhancing the consumer register

Data were linked from two main sources, with the aim of acquiring a representative register of the UK population at large: the CACI Consumer Register and birth certificate data from the UK Office for National Statistics. The age-sex structure as recorded in the 2011 Census was also used in order to standardise the distribution recorded from the combined name data.

CACI Ltd. (London, UK) provided the original derived data product called Monica, which contains details of age and gender distributions associated with different given names amongst adults. The data were extracted from credit card applications, and pertained to a total of 7,085,617 individuals, who were bearers of over 21,000 individual names (multi-gendered names are counted twice). For reasons of disclosure control only names with a sample size of at least 10 were included in the dataset. However, the dataset had two key limitations because of the nature of its remit, viz. credit card approvals. First, no applicants were aged under 18. Second, certain age groups were underrepresented relative to their known frequencies in the UK, particularly those aged 18 and 19. There is also a possible limitation that some names which are more prevalent amongst more deprived households may be slightly underrepresented because of socio-economic inequalities within the credit card market.

To establish a more representative and inclusive age structure, birth certificate data were obtained from the Office for National Statistics (ONS). Since 1996 the ONS has released data detailing the frequency of births registered in each given name. As of 2012 the birth certificate data accounted for a cumulative population group of 10,412,724 individuals, excluding individuals with names with a frequency of 3 or less in any given year in England and Wales. While these data cannot account for children and teenagers born outside of England and Wales, it nevertheless represents a substantial proportion of the overall UK population for this age range. Combined, the two datasets represent over 32,000 unique forenames, although only 1441 of these have a collective sample size greater than 1000 individuals.

The forenames were not recoded into their most common variants: for instance Matt and Matthew were kept as separate and distinguishable names. Parental choice of shortened baby names peaked in popularity about 25 years ago. Indeed the Consumer Register data from 2011 revealed that Output Areas with higher proportions of young adults had the highest concentrations of shortened names such as Tom, Alex, Nick, Joe and Sam.

Our final names database has been aggregated into 5 year age bands, consistent with the CACI Monica classification. The datasets were reweighted to account for the uneven sample sizes based on their penetration of the UK population. The primary aim was to develop a model of age and gender distributions which was reflective for the entire UK population across each age group. Age and gender distributions were calculated and were found to be very similar to the overall age distribution of the UK, with the exception of ages 18–19 (when looking at the data within the 15–19 age group) and a very slight underrepresentation of elderly age groups. It was therefore decided to reweight the entire dataset by the UK population using official statistics from the 2011 Census using the same age bands.

4. Key trends in naming practices

The combined names database reveals that there are more unique female names than individual male names in the UK. The top 10 most common male names account for 25.13% of the male population, whilst the top 10 female names account for 13.07% of the female population. The frequency distributions of names for both genders are very positively skewed. Lorenz Curves for the frequencies of male and female names revealed that a minority of names represent a majority of the population (Fig. 1). Although female names are slightly more evenly distributed.

The most common forenames for both genders are shown in Table 2. The data reveal that John is the most common male name, and Margaret is the most common female name. Other research from the ONS confirms that these names were the most common baby names in the earlier years of the 20th Century (Matheson and Babb, 2002), and it is therefore likely that the bearers of these names from the 2011 Consumer Register derive from older age cohorts.

The data also reveal that the choice of baby name has diversified over time. For instance the most popular baby names for the youngest age band, Oliver and Olivia, represent 2.2% and 1.6% of all persons from the 0 to 5 age band for each gender respectively. By contrast, the most popular names for the eldest age band, Margaret and John, represent 5.6% and 8.3% of persons for each gender.

The name data have confirmed that popular naming trends have fluctuated over time, the most popular names for each age band are presented in Table 3. Consequently, certain names can be generally associated with particular age groups. This can be easily demonstrated by taking the average age for each given name. Using this technique, the ten oldest and youngest names for both genders are presented in Table 4.

However, the average age alone is not an appropriate indicator for understanding the typical age structure for every name. Many name distributions manifest a peak in popularity which gradually diminished over time. For plenty of these names, the rise to popularity was gradual. Yet overall they do not quite share normal distributions across ages due to the age structure of the population and all frequencies must be truncated at age zero. However, generally trends in names seem to fluctuate gradually over time due to shifts in baby name popularity and changes in migration, and most have a distinctive peak.

Some forenames have been more resilient to fluctuations in popularity over time, and these are therefore well represented across a wide range of age groups today. For instance the aggregated age structure of those with the names Patrick, James, Catherine, Ruth and Robert vary little from the national average. When comparing the relative proportion of the age distributions from the names to the national population, Patrick, George and Edward share the smallest standard deviations (0.01). In contrast

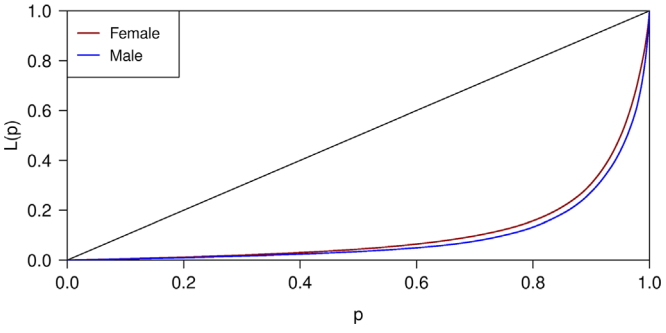

Fig. 1. Lorenz curves demonstrating the frequency distributions of male and female names. Only names with a projected frequency of 1000 were included in this visualisation.

Table 2
The most common forenames, as identified in the combined database (England and Wales weighting).

	Female		Male	
Rank	Name	Estimated number	Name	Estimated number
1	MARGARET	555,000	JOHN	1,003,500
2	SUSAN	544,600	DAVID	969,000
3	SARAH	425,600	JAMES	614,600
4	ELIZABETH	363,200	MICHAEL	612,800
5	PATRICIA	350,200	PAUL	546,700
6	MARY	332,500	ROBERT	482,800
7	CHRISTINE	321,400	PETER	480,600
8	JULIE	314,600	ANDREW	435,700
9	KAREN	313,700	WILLIAM	415,300
10	LINDA	298,200	MARK	370,700

Table 3
The most common name for each age band.

Age group	Female	Male
0–4	OLIVIA	OLIVER
5–9	EMILY	JACK
10–14	CHLOE	JACK
15–19	SARAH	JAMES
20–24	SARAH	JAMES
25–29	SARAH	DAVID
30–34	SARAH	DAVID
35–39	SARAH	PAUL
40–44	KAREN	PAUL
45–49	JULIE	DAVID
50–54	SUSAN	DAVID
55–59	SUSAN	DAVID
60–64	SUSAN	JOHN
65–69	MARGARET	JOHN
70–74	MARGARET	JOHN
75–79	MARGARET	JOHN
80–84	MARGARET	JOHN
85+	MARGARET	JOHN

Table 4
The oldest and youngest names in the UK (excluding names with a projected population under 10,000).

	Female		Male	
Rank	**Oldest**	**Youngest**	**Oldest**	**Youngest**
1	DORIS	SIENNA	CYRIL	RILEY
2	GLADYS	AVA	HERBERT	JAYDEN
3	ETHEL	EVIE	REGINALD	LOGAN
4	EDNA	SUMMER	ERNEST	FINLEY
5	WINIFRED	LACEY	HAROLD	NOAH
6	HILDA	SCARLETT	WALTER	ALFIE
7	BETTY	GRACIE	RONALD	LUCAS
8	MURIEL	MADDISON	LEONARD	HARLEY
9	PEGGY	FREYA	NORMAN	FREDDIE
10	VERA	MADISON	DONALD	KIAN

Oscar, Archie, Jordan and Shannon share deviations above 0.11 – all four have very young skewed age distributions.

Certain names exert bimodal age distributions, perhaps reflecting an inter-generational popularity. For instance the name Guy is most common amongst men in their mid to late forties, and second most popular amongst those in their mid-twenties, a gap of just over 20 years. Madeline is common amongst older children and also those in their 60 s too. There has also been a resurgence in popularity of some names which were also popular in the 1930s as revealed by the data. Notable examples include Clara, Rose, Sidney and Henry.

It is not uncommon for influential celebrities' baby names to

gain popularity: for example, over 96% of persons named Brooklyn in the UK are younger than Brooklyn Beckham. However, this is not always the case, for instance there are only 10 instances of the name Apple in our dataset, the moniker given to Chris Martin and Gwyneth Paltrow's first child which caught the attention of the media. Celebrities' own names may also influence baby naming: 63% of those named Rihanna were under five years of age as of 2012, and there are no records of this name in the original Monica file which only represents those aged 18 and older. 46% of those named Beyoncé are in the 5 to 9 age group. The name Rod is most

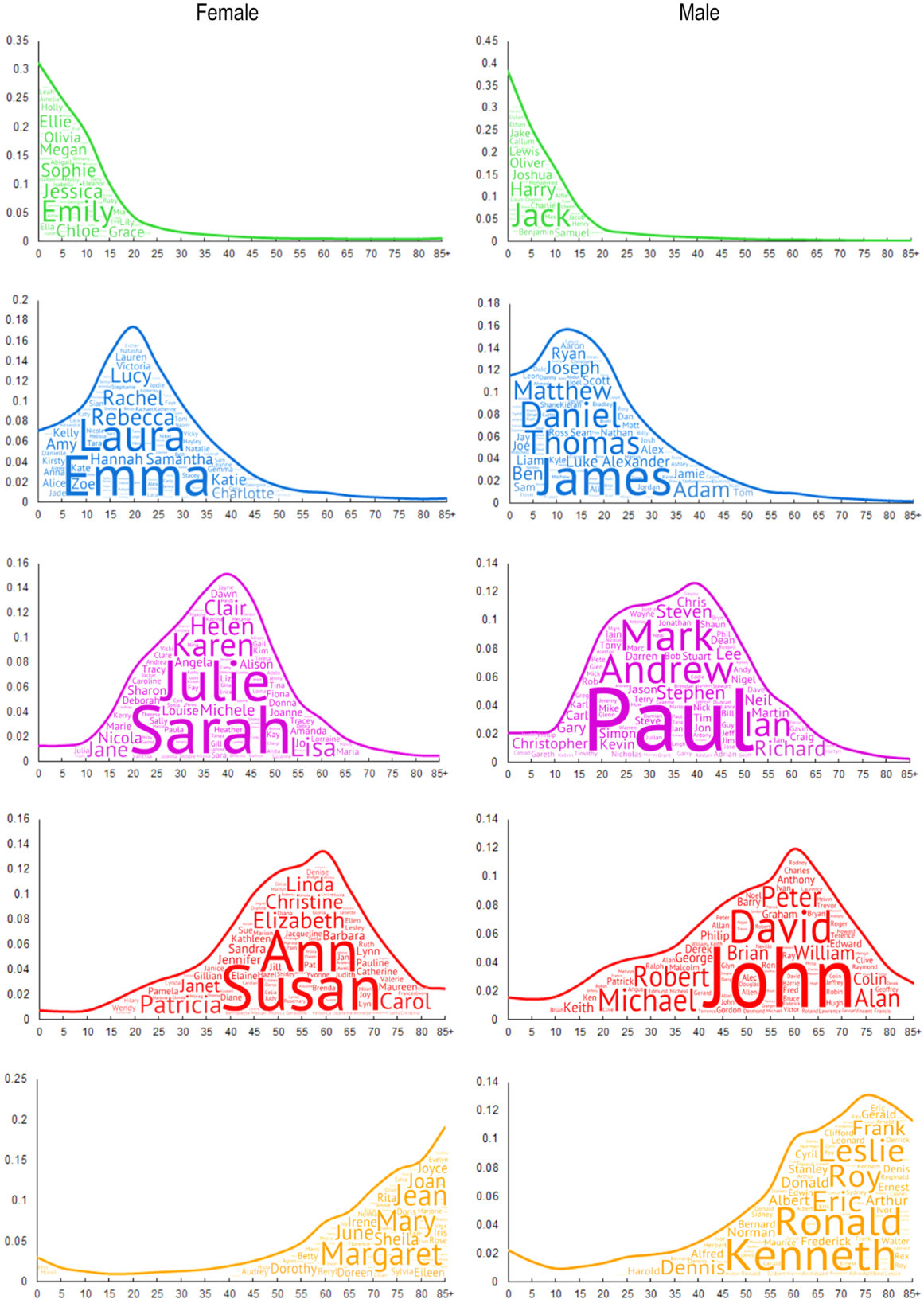

Fig. 2. The age distributions of the five name clusters based on five year age bands for each gender. From top to bottom; Youngest, Young, Medium, Old, Oldest.

Table 5
The names with the oldest mean ages in Great Britain (Excluding those with a projected population below 10,000).

	Female		Male	
Rank	Name	Estimated age	Name	Estimated age
1	DORIS	78.35	CYRIL	73.55
2	GLADYS	77.99	HERBERT	72.10
3	ETHEL	77.74	REGINALD	71.90
4	EDNA	76.92	ERNEST	70.84
5	WINIFRED	76.75	HAROLD	70.13
6	HILDA	76.64	WALTER	68.85
7	BETTY	76.46	RONALD	68.61
8	MURIEL	75.80	LEONARD	67.24
9	PEGGY	75.75	NORMAN	66.95
10	VERA	75.58	DONALD	66.02

common amongst those in their late thirties to their forties, perhaps this was driven by the influence of Rod Stewart's popularity, whose career peaked in the late 60 s and early 70 s.

4.1. Grouping names

To achieve an overview of how forenames can be generalised based on their standardised age distributions, the names have been clustered by running a k-means clustering algorithm for each gender. The k-means algorithm is an interactive process which clusters data by allocating and reallocating observations to their nearest cluster centroid in a multidimensional space as determined by the variables. The algorithm attempts to minimise the average distance of each observation to its nearest centroid with each iteration by updating (and therefore moving) the centroids based on their existing allocation of observations and then reallocating the data. This process continues until an optimum solution has been achieved and the centroids can no longer be moved (Harris et al., 2005). The number of clusters are specified by the researcher and their centroids are initially randomly located when the algorithm commences. Only names with a frequency of 1000 or more in the original data were included in this analysis and the variables inputted were the proportion of their populations within each 5 year age band. The classification produced five groups for both genders and these have been labelled; youngest, young, middle, old, oldest based on their average ages. Each group generally represents a 20 year generation of popularity. The results of the classification have been visualised as a series of frequency probability histograms. The size (in area) of each name within the histograms corresponds with its projected population size in Fig. 2.

The classification groups are reasonably well balanced in terms of both the number of names and their estimated populations. However, the youngest name group contains the most unique monikers but is the second smallest group for both genders in terms of expected population of the bearers: this reflects the increased the diversity of names over time, with the oldest groups being the smallest in both the frequency of individual names and total population.

There are some notable distinctions between the male and female classifications in that the oldest group for the male population is smaller in size and has a much flatter distribution than its female equivalent. The youngest age group is well represented by contemporary popular baby names such as Jack, Harry, Emily and Jessica. Bearers of these names typically average 10.4 years of age for males, and 12.5 for females, and 75% of females and 84% of males are under the age of 15 on average for each name within the groups. There is also a higher proportion of names imported from abroad, which can be accounted for by the higher fertility rates amongst their parents' cohort (Zumpe et al., 2012). Mohammed is the 11th most popular male name in the Young group.

The young, medium and old groups all display relatively normal histograms suggesting that the popularity of such names have gradually risen and fallen again. On average 53% of the young group for both genders are aged between 10 and 30, the male equivalent being slightly younger on average. The most popular names in these groups are Emma and Laura for females and James, Thomas and Daniel for males. The medium groups are represented by large numbers of adults aged between the ages of 20 and 50, with peaks in the forties. Names from this generation are well represented by Sarah, Julie and Karen for females and Paul, Andrew and Mark for males.

The old group is most prevalent amongst persons aged between 50 and 70 years of age. Although these names have declined in popularity, this phenomenon has been quite gradual and subsequently about 4% of bearers of these names are aged 25–30. The male names assigned to this category have remained more prevalent in recent years than their female counterparts. This is probably because of the presence of inter-generational use of recurring family forenames, some of which are biblical in origin (such as John, David, Peter and Michael).

The oldest group are very advanced in years. Cyril, the oldest male name in our dataset, has an average age of 73.5 and a median age of 76.5, whilst the oldest female name, Doris, has an average age of 78.4. The ten names with the oldest average ages for each gender are presented in Table 5. Interestingly, the data also identify the recent resurgence in popularity of Victorian names (West, 2012), and there is a subtle rise in the frequency of these names amongst the youngest age cohort too.

5. Modelling demographic structures from names

As a test of concept, the names classification has been used to estimate the age structure of georeferenced Twitter users in London using a sample of 2.5 million Tweets from September 2012 to March 2013, from 129,400 individual users. Twitter data has been widely utilised by researchers and marketers to gain an understanding about people and their opinions (Williams et al., 2013). However, the demographic characteristics of Twitter users are poorly understood as no such information is recorded by Twitter. For this study, users' forenames were extracted from the user names using an approach outlined in Longley et al. (2015). To gain an understanding how the demographics of users may vary across space, the data were partitioned for four distinctive places: an entertainments venue (the O2 Arena), a football stadium (the Emirates Stadium), a business district (Canary Wharf) and a shopping centre (Westfield, Stratford). The modelled age pyramids are displayed in Fig. 3, where the grey lines represent the average distribution across all the Tweets in London from the original sample.

Inference of age from the database identifies the football stadium and business district as places where there are greater proportions of male Tweeters. It also identifies younger age distributions at the shopping centre and entertainments venue. Both would appear to be logical conclusions based on the known activities of these places. Of course, it is very unlikely that persons under the age of 5 are Twitter users. The presence of very young names in the Twitter dataset, which is particularly evident at the football stadium, can be largely accounted for by the use of informal spellings of names. However, future analysis could refit the names dataset to accommodate a minimum age cut-off if the data we are trying to model are known to only represent the adult population. In summary, whilst the findings for the four sample locations are unsurprising, it indicates that the analysis can be conducted in other locations where the demographic composition

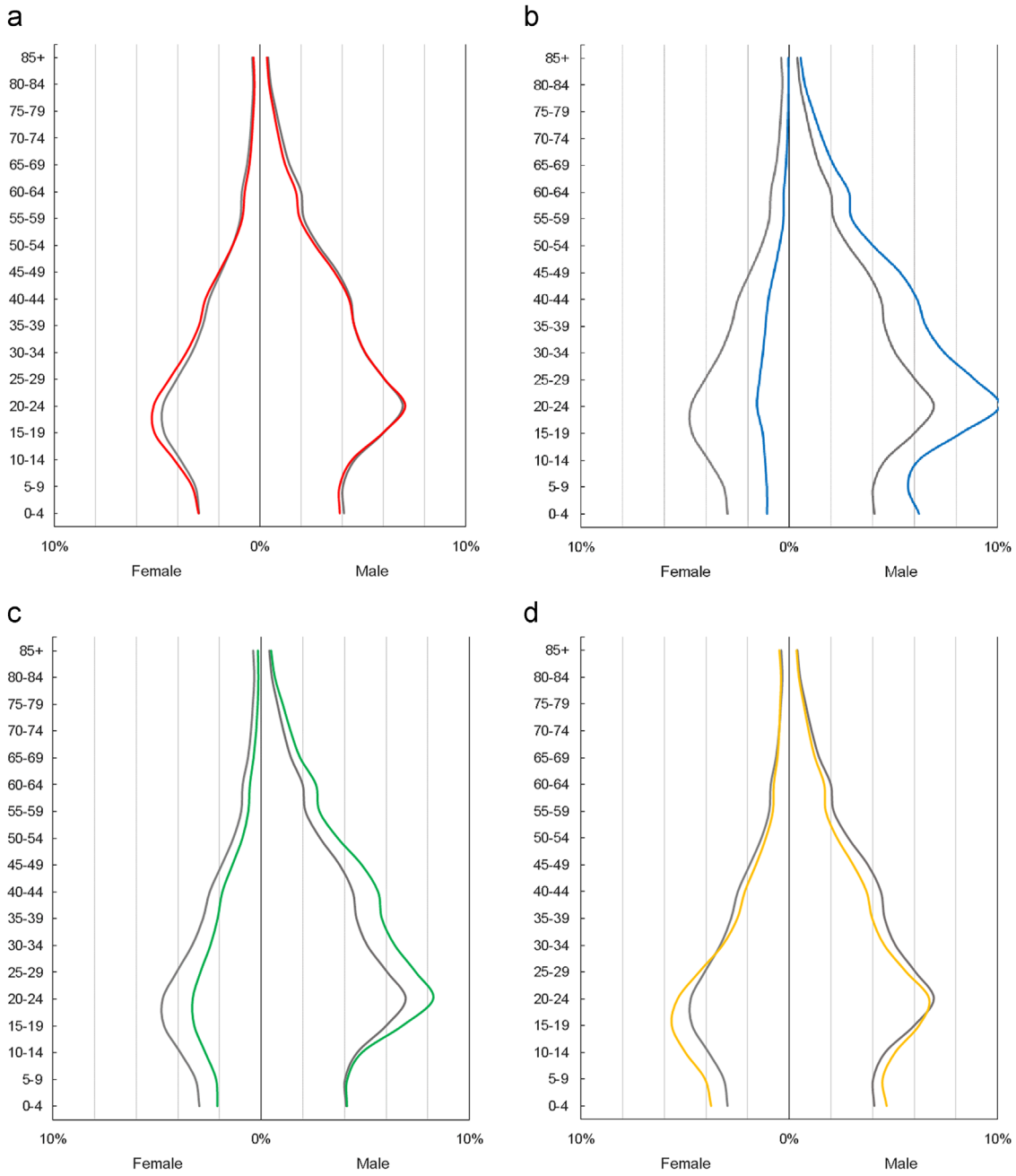

Fig. 3. Name inferred demographic structure of geotagged Twitter users at four sites in London. a) The O2 Arena, b) The Emirates stadium, c) Canary Wharf and d) Westfield Stratford.

are not known. It also indicates that gender and ages can be inferred from names from other datasets too.

6. Conclusions

This paper has demonstrated that it is feasible to identify the age-gender distributions of forenames and thence to ascribe demographic characteristics to data where such information is not otherwise available. This approach makes it possible to unshackle geodemographic analysis from an exclusive preoccupation with the geography of night-time residence, and harness value from individual level data (Longley and Adnan, 2016). The modelled name data provide a suitable means of estimating age and gender distributions from British forenames because of trends in the popularity of baby names. Such data could supplement the analysis of population records such as customer datasets, which inherently lack demographic characteristics. The inferential procedure is of course by no means perfect, not least because of the incompleteness of the adult population that are included in the market data and the presence of young people who were not born in England and Wales. In addition, birth certificate records could not be obtained for Scotland or Northern Ireland, and the database required further reweighting to account for uneven sample sizes between the two data sources. It is also important to consider the limitations of individual level data in the consumer context. A person may not always be shopping for his or herself exclusively. For example, a grocery shopping trip could be undertaken by an

individual for their entire household.

However, such an approach is a viable means of assigning characteristics to individuals in customer databases which contain few if any demographic attributes but do include names. The database can be used to ascribe the probability of each individual falling into each age band and gender. Of course, each of the names is uniquely distributed across ages and genders and some may be more uniformly distributed than others, making them less effective discriminators. However, most names have been found to be broadly representative of particular age groups. Names are also particularly successful as a means of estimating gender as the vast majority of names are not unisex. Previous research has identified that consumer's product preferences vary considerably by age and gender. Therefore harnessing demographic attributes of consumers at the individual level is of great benefit to retailers, particularly those which are able to invest in micro-level targeted marketing strategies. In this context, the analysis of forenames can empower retailers to take advantage of better insight from their data to inform their future marketing and planning decisions.

Acknowledgements

This research was funded by Economic and Social Research Council grants ES/L011840/1 (Consumer Data Research Centre) and ES/L013800/1 (The Analysis of Names from the 2011 Census of Population).

References

Bakewell, C., Mitchell, V.W., 2003. Generation Y female consumer decision-making styles. Int. J. Retail Distrib. Manag. 31 (2), 95–106.
Berger, J., Mens, G., 2009. How adoption speed affects the abandonment of cultural tastes. Proc. Natl. Acad. Sci. 106 (20), 8146–8150.
Blázquez, M., 2014. Fashion shopping in multichannel retail: the role of technology in enhancing the customer experience. Int. J. Electron. Commer. 18 (4), 97–116.
Baudrillard, J., 1998. The Consumer Society: Myths and Structures. Sage, London.
Carpenter, J.M., Moore, M., 2006. Consumer demographics, store attributes, and retail format choice in the US grocery market. Int. J. Retail Distrib. Manag. 34 (6), 434–452.
Campbell, C., 1997. Shopping, pleasure and the sex war. In: Falk, P., Campbell, C. (Eds.), The Shopping Experience. Sage, London, pp. 166–175.
Dittmar, H., Long, K., Meek, R., 2004. Buying on the internet: gender differences in on-line and conventional shopping motivations. Sex Roles 50 (5), 423–444.
Erwin, P., Calev, A., 1984. The influence of Christian name stereotypes on the marking of children's essays. Br. J. Educ. Psychol. 54, 223–227.
Finch, J., 2008. Naming names: kinship, individuality and personal names. Sociology 42 (4), 709–725.
Galbi, D.A., 2002. Long-term trends in personal given name frequencies in the UK. Names 50 (4), 275–288.
Gale, C.G., 2014. Creating an Open Geodemographic Classification Using the UK Census of the Population. Doctoral thesis. University College London, Department of Geography, London.
Gureckis, T.M., Goldstone, R.L., 2009. How you named your child: understanding the relationship between individual decision making and collective outcomes. Top. Cognit. Sci. 1 (4), 651–674.
Harari, H., McDavid, J.W., 1973. Name stereotypes and teachers' expectations. J. Educ. Psychol. 65 (2), 222–225.
Harmon, H.H., Webster, R.L., Weyenberg, S., 1999. Marketing medium impact: differences between baby boomers and Generation Xers in their information search in a variety of purchase decision situations. J. Mark. Commun. 5 (1), 29–38.
Harris, R., Sleight, P., Webber, R., 2005. Geodemographics, GIS and Neighbourhood Targeting. John Wiley and Sons, Chichester.
Hensher, D.A., Johnson, L.W., 1981. Applied Discrete Choice Modelling. Croom Helm, Beckenham.
Kalyanam, K., Putler, D.S., 1997. Incorporating demographic variables in brand choice models: an indivisible alternatives framework. Mark. Sci. 16 (2), 166–181.
Kohijoki, A., Marjanen, H., 2012. The effect of age on shopping orientation—choice orientation types of the ageing shoppers. J. Retail. Consum. Serv. 2 (2), 165–172.
Lieberson, S., 2000. A matter of taste: How names, fashions, and culture changes. Yale University Press, New Haven.
Longley, P.A., Adnan, M., 2016. Geo-temporal Twitter demographics. Int. J. Geogr. Inf. Sci. 30 (2), 369–389. http://dx.doi.org/10.1080/13658816.2015.1089441.
Longley, P.A., Adnan, M., Lansley, G., 2015. The geotemporal demographics of Twitter usage. Environ. Plan. A 47 (2), 465–484.
Lord, E., 2002. Given names and inheritance. In: Postles, D. (Ed.), Naming, Society and Regional Identity 169 – 192. Leopard's Head Press, Oxford.
Lunt, P.K., Livingstone, S.M., 1992. Mass consumption and personal identity. Open University Press, Buckingham.
McDonald, W.J., 1995. Time use in shopping: the role of personal characteristics. J. Retail. 70 (4), 345–365.
Mateos, P., 2007. A review of name-based ethnicity classification methods and their potential in population studies. Popul. Space Place 13, 243–263.
Mateos, P., Longley, P.A., O'Sullivan, D., 2011. Ethnicity and population structure in personal naming networks. PLoS One 6 (9), e22943 1–12.
Matheson, J., Babb, P. (Eds.), 2002. Social Trends 32. The Stationery Office, London.
Matheson, J., Summerfield, C. (Eds.), 2002. Social Trends 30. The Stationery Office, London.
Mitchell, V.W., McGoldrick, P.J., 1994. The role of geodemographics in segmenting and targeting consumer markets: a Delphi study. Eur. J. Mark. 28 (5), 54–72.
O'Malley, L., Patterson, M., Evans, M., 1997. Retailer use of geodemographic and other data sources: an empirical investigation. Int. J. Retail Distrib. Manag. 25 (6), 188–196.
ONS, 2015. Output Area (OA). The Office for National Statistics website. Online: ⟨www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area–oas-/index.html⟩ (accessed 05.03.15).
Openshaw, S., 1984. Ecological fallacies and the analysis of area census data. Environ. Plan. A 16, 17–31.
Pentecost, R., Lynda, A., 2010. Fashion retailing and the bottom line: the effects of generational cohorts, gender, fashion fanship, attitudes and impulse buying on fashion expenditure. J. Retail. Consum. Serv. 17 (1), 43–52.
Reynolds, F.D., Wells, W.D., 1977. Consumer Behaviour. McGraw Hill, New York.
Rodgers, S., Harris, M., 2003. A gender and e-commerce: an exploratory study. J. Advert. Res. 43 (3), 322–329.
Scanlon, J. (Ed.), 2000. The Gender and Consumer Culture Reader. New York University Press, New York; London.
Scharf, T., 2005. Recruiting older participants: lessons from deprived neighbourhoods. In: Holland, C. (Ed.), Recruitment and Sampling: Qualitative Research With Older People. Centre for Policy on Ageing, London, pp. 29–43.
Smith-Bannister, S., 1997. Names and Naming Patterns in England. 1538 – 1700. Oxford University Press, Oxford.
Sorce, P., Perotti, V., Widrick, S., 2005. Attitude and age differences in online buying. Int. J. Retail Distrib. Manag. 33 (2), 122–132.
Wardle, J., Haase, A.M., Steptoe, A., Nillapun, M., Jonwutiwes, K., Bellisle, F., 2004. Gender differences in food choice: the contribution of health beliefs and dieting. Ann. Behav. Med. 27, 107–116.
West, E., 2012. Say hi to Ethel Mary: Victorian names are back in fashion. The Telegraph Blogs. ⟨http://blogs.telegraph.co.uk/news/edwest/100127208/say-hi-to-ethel-mary-victorian-names-are-back-in-fashion/⟩.
Williams, S.A., Terras, M., Warwick, C., 2013. What people study when they study Twitter: classifying Twitter related academic papers. J. Doc. 69, 3.
Xi, N., Zhang, Z., Zhang, Y., Ge, Z., She, Le, Zhang, k, 2014. Cultural evolution: the case of babies' first names. Phys. A: Stat. Mech. Appl. 406, 139–144.
Zumpe, J., Dormon, O., Jefferies, J., 2012. Childbearing Among UK Born and Non-UK Born Women Living in the UK. The Office for National Statistics ⟨http://www.ons.gov.uk/ons/dcp171766_283876.pdf⟩.