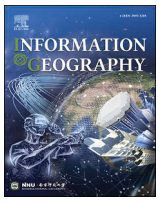
Information Geography 1 (2025) 100023

Contents lists available at ScienceDirect

Information Geography

journal homepage: www.sciencedirect.com/journal/information-geography



Research Article

From names to numbers: Modelling age and gender profiles from consumer data

Maurizio Gibin [a], Justin van Dijk [a], Zi Ye [b], Paul Longley [c,*]

[a] University College London, Geography, United Kingdom
[b] University of Liverpool, Geography and Planning, United Kingdom
[c] University College London, United Kingdom

ARTICLE INFO

Keywords:
Age demographics
Consumer registers
UK census of population
Lifestyles survey
Baby names

ABSTRACT

It is axiomatic to Information Geography that, wherever possible, data about human subjects should be created and maintained at the level of the individual. This paper develops and evaluates an innovative approach to inferring the ages of individuals from their given and family names. We use a major UK consumer lifestyles survey alongside baby name statistics to establish the age distributions associated with a comprehensive range of given names. We also use the mix of adult given names within different types of households to refine our age estimates for specific individuals. We evaluate the accuracy of these estimation techniques with respect to (a) specific respondents to a lifestyles survey and (b) UK Census small area estimates. We describe how this approach can be used to ascertain the representativeness of new sources of data and suggest further ways in which the methods might be refined using other contextual information.

1. Background

Social science interest in information geography has been fuelled by vast recent increases in the amount of data that are collected about citizens today. However, because ever increasing proportions of these data are collected through consumer transactions, rather than conventional government social surveys, it is no longer the case that the terms 'data' and 'information' can be thought of as near synonyms (Longley and Chen, 2025). 'Smart' data collected as a result of human interactions with digital devices during provisioning of goods or services are a by-product of these transactions rather than any integral part of research design or social investigation. Turning data into useable information frequently requires consideration of the population from which typically self-selecting data subjects are drawn and diagnosis of the sources and operation of bias when comparing the elements of the represented population with those that are absent. This typically requires careful triangulation with conventional statistical sources that, although typically less rich in detail and less frequently collected, adhere to a conventional research design with sampled elements having known and prespecified probabilities of inclusion.

With these crucial provisos, new sources of smart data can greatly enrich information geographies of the socioeconomic realm, with frequent updates. There is, however, a further caveat that few organisations that collect smart data have monopoly positions in the delivery of a full range of digital or physical services and, as such, representation of the human self is 'Balkanised' into shards of data rather than the holistic ranges of characteristics that are required to support most social investigations (Goodchild, 2022). Thus, while consumer and administrative data have become increasingly important for studying processes of population change, the requirement to provide a comprehensive range of characteristics to support social investigations requires linkage of different data sources or modelling of missing characteristics. The inherent ambiguities of linking data that pertain to aggregations of individuals present in areas creates possible issues of ecological fallacy in geographic analysis (Openshaw, 1984). Lined smart data representations of individuals are thus best grounded at the level of the human individual.

An example of UK data infrastructure derived from various consumer and administrative sources is the Linked Consumer Registers (LCRs: Lansley et al., 2019). This annual series of UK adult names and addresses is geographically referenced, maintained and updated at the level of the individual citizen. It is primarily built from electoral registers but then supplemented with consumer datasets to capture individuals not registered to vote or otherwise absent from the public version of the register. For every year between 1997 and the present day, the LCRs provide georeferenced individuals' names and addresses

* Corresponding author.
E-mail address: p.longley@ucl.ac.uk (P. Longley).

https://doi.org/10.1016/j.infgeo.2025.100023
Received 16 April 2025; Received in revised form 23 May 2025; Accepted 13 June 2025
Available online 20 June 2025

that enable a highly disaggregated and frequently updated representation of population size and household structure. The data are collected and maintained by the Geographic Data Service (GeoDS: https://data.geods.ac.uk/dataset/linked-consumer-registers), which is part of the Smart Data Research UK initiative (www.sdruk.ukri.org/about-smart-data-research/), A principal motivation for the creation and maintenance of these data is that modelled characteristics of individuals and households are best grounded at the level of the individual and that frequently updated highly disaggregate estimates offer new ways of measuring and modelling population characteristics for any convenient (and non-disclosive) geographic aggregation. Derivative 'research ready' (Longley et al., 2024) datasets include small area modelled ethnicity proportions (https://data.geods.ac.uk/dataset/modelled-ethnicity-proportions-lsoa-geography), estimates of distances of residential moves (https://data.geods.ac.uk/dataset/distances-of-residential-moves-dorm-index-lad-geography) and neighbourhood measures of the changes in neighbourhood living conditions that accompany residential moves (https://data.geods.ac.uk/dataset/residential-mobility-and-deprivation-rmd-index-lad-geography). These research ready datasets are made available for bona fide public good research upon successful application to the GeoDS.

For the principal year used for this study, the 2011 LCR lists similar numbers of adult residents throughout the UK to the Census of Population conducted in that year. The strength of the LCRs lies in providing detailed individual-level data that, with disclosure control safeguards, can be linked and aggregated to any desired geography. For example, Van Dijk et al. (2021) use the LCRs to develop longitudinal analysis of intra-national migration and residential mobility; and Kandt et al. (2020) couple historic censuses with contemporary LCRs to investigate population change over several generations.

While the multiple linked data sources that make up LCRs can be used to study changes in the presence or absence of individuals and households using highly granular geographies, individual and household level characteristics such as gender, age and ethnicity are not available. Previous research has begun to address these issues through estimating individual ethnicity characteristics by examining personal names: research conducted in partnership with the Office for National Statistics (ONS: Kandt and Longley, 2018; Lan and Longley, 2023) has developed individual-level name-based ethnicity classifications that have been used to develop scale-free analysis of ethnic segregation (Lan et al., 2020) and to chart inter-generational social and spatial inequalities of outcome (Longley et al., 2018). Lansley and Longley (2016) pilot an approach to investigate the gender and general age characteristics associated with some forenames in Britain, exploiting variations in the popularity of different given names over time.

This study explores aggregate generalizations of individuals' ages in relation to their household structures, and validates its findings. By extending and updating previous work on name-based age and gender profiling, we develop a contemporary classification for 19,000 given names. We construct new household models of joint age distributions to improve age estimates, and then link a 2022 lifestyles survey to validate the age models. The lifestyles survey, conducted by PDV Ltd., is a commercial dataset developed for marketing and research purposes in the United Kingdom, and is also made available for academic use by the Geographic Data Service (GeoDS: data.geods.ac.uk/dataset/pdv-consumer-lifestyle-surveys). The dataset contains over 17 million individual records, collected from 2008 to 2021, encompassing adult respondents' names, addresses, birth dates, gender, and household characteristics. The PDV dataset is a large and rolling private sector survey that, with appropriate user consents, is used for marketing purposes by a large and varied private sector client base. Participation in the survey is voluntary and through various online and other channels. As such the survey exhibits some bias, most evidently for present purposes in the age and gender distribution of respondents which are disproportionately female and over 30 years of age. There is also bias in reported housing tenure, with mortgage holders significantly over-represented (39.25%) and outright homeowners strongly under-represented (10.10%), compared to Census 2021 estimates of 28.74% and 32.83% respectively. In terms of housing type, the PDV survey under-represents residents of flats and detached houses while slightly over-representing those in semi-detached homes. Marital status, however, is broadly in line with national distributions, and there is no reason to anticipate that the names of those included deviate from those of the population at large.

It is axiomatic to Information Geography that, wherever possible, data about human subjects should be created and maintained at the level of the individual, in order to anticipate the risks of ecological fallacy when such data are analysed. In this context, this study is also new and novel in three respects. Firstly, we not only employ individual probabilities to assign ages to individuals but also use the joint probabilities of multiple adult household members to refine them. Secondly, we compare our estimates against the ONS 2011 Census single-year reported age frequencies for Lower layer Super Output Areas (LSOAs). Finally, we validate the final estimated individual ages by using data from individuals who participated in the PDV survey.

2. Development and application of an age and gender predictive model

Our approach is built around two pillars, each of which serves a unique purpose within the broader scope of information geography. Firstly, in seeking to replicate 2011 Census estimates, we develop a model which enables us to use annual LCRs to determine the ages of individuals in any year of the LCRs pre- or post-2011. This approach enables micro-demographic analysis of changes in the ages of neighbourhood residents at any convenient scale of analysis. Secondly, and of potentially greater significance, is the development of a more general predictive model that can be applied to any other UK address list which includes household composition. The flexibility of this model, and its general applicability to other datasets, makes it a high-value output for demographic research, including the triangulation of smart data with conventional statistical or administrative sources in order to establish their provenance. As such, it can be seen as a contribution to a developing paradigm in which intelligent data services distil 'research ready' extracts from assemblages of smart or administrative data in ways that are efficient, effective and safe to use across the research community (Longley and Chen, 2025; Longley et al., 2024; McGrath-Lone et al., 2022).

The core of the model lies in predicting the age and gender of individuals based on the available PDV survey data. To enhance the accuracy of these predictions, we propose that gender-specific age estimates are adjusted using weights derived from a comparison between ages recorded in the lifestyles survey and the 2011 Census of Population. In developing a joint model for households with at least two members, it is suggested that the model be weighted by the statistical moments of the component name distributions. For example, the age predictions for names with different distributions, such as John (platykurtic, negatively skewed, high standard deviation) versus Kylie (leptokurtic, positively skewed, low standard deviation), should account for these variations. For common forenames, we introduce weighting using the reciprocal of the variance as an initial step. The model also addresses the challenge of handling rarer names that may not exhibit continuous distributions.

We are guided in this endeavour by the findings of past studies in this underdeveloped area of research. Our research design is formative, entailing exploration of different measures of central tendency and

permutation of joint name distributions to suggest a best global fit against exacting performance metrics.

3. Data used for age and gender estimation

The age and gender estimations in this study rely on multiple datasets that provide a comprehensive view of individuals' demographics, particularly names, ages, and household structures. The primary datasets used are the 2011 LCR and the 2023 PDV Consumer Lifestyle Survey. While the LCRs do not include age or gender, they provide a strong backbone for individual level linkage with the PDV survey responses, specifically of name, gender and date of birth. For the purposes of this study, we select the 2011 LCR because aggregated estimates are directly comparable with the UK Census of Population, enabling evaluation of age estimates against a well-established demographic dataset (Van Dijk et al., 2021). Inclusion of most all-adult household members in the 2011 LCR enables the development of joint age distribution models for households (see Fig. 1).

Given the PDV survey's restriction to adults, additions were required to complete forename age distributions with younger age cohorts. UK-wide ONS counts of forenames of newborns were compiled for each year since 1997, and were integrated with the PDV data to fill in gaps for individuals under 27 (see Fig. 1). In addition to providing estimates for minors, these counts were used to overwrite PDV data for young adults, as participation rates in the Survey are observed to be low in the youngest adult age groups (Fig. 2). Gender was assigned to the category that accounted for more than 50% of bearers using the combined and weighted PDV and ONS datasets. We next calculated the total 27+ population from the 2021/2 Census and devised the weights required to gross each annual PDV or ONS baby names sample to its corresponding population size estimate (Fig. 3). These grossing factors were applied to each given name when estimating the age distribution of bearers of that name. An illustration of the age distribution of bearers of the name Paul in 2023 is shown in Fig. 4.

Our inference of ages from given names in 2011 relies upon changes in the popularity of names over time, observed in 2023 – by which time the extents of the older age cohorts in Fig. 3 had been eroded by mortality. Accordingly, we utilized ONS Life Tables to adjust the observed age distributions to account for the higher gender-specific mortality rates in successively higher annual age bands. For each forename, we estimate the number of people aged *i* in year *t-1* as:

$$P_{t-1,i-1} = \frac{P_{t,i}}{1 - d_{t-1,i-1}} \quad (1)$$

where:

$P_{t,i}$ is the population aged *i* in year *t* (e.g. 2023).

$d_{t-1,i-1}$ is the death rate for each age *i-1* during year *t−1* (e.g. 2022) to *t*.

The formula assumes net migration to be zero and uses the ONS Life tables (Office for National Statistics, 2024) to estimate the fraction of each 2011 age cohort that had deceased. We then reweighted each Age and Gender by Forename (AGF) distribution by the difference between this figure and the count of survivors in each single age. This process enables estimation of the complete probability density distributions for any forename in any chosen year – specifically in our research, to enable comparison of age estimates with 2011 Census data. The intermediate AGF dataset contained age frequency distributions for more than 19,000 forenames.

Fig. 5 compares the age pyramid estimates from the 2011 UK Census with the mortality adjusted 2011 intermediate AGF data. Age-specific weights were added to the grossing factors used in Fig. 3 to reconcile the combined AGF distribution with 2011 Census estimates. These weights were then applied to each occurrence of a given name in every age cohort in which it occurred. Fig. 6 shows the resulting age distribution for the name Paul using the final 2011 AGF model.

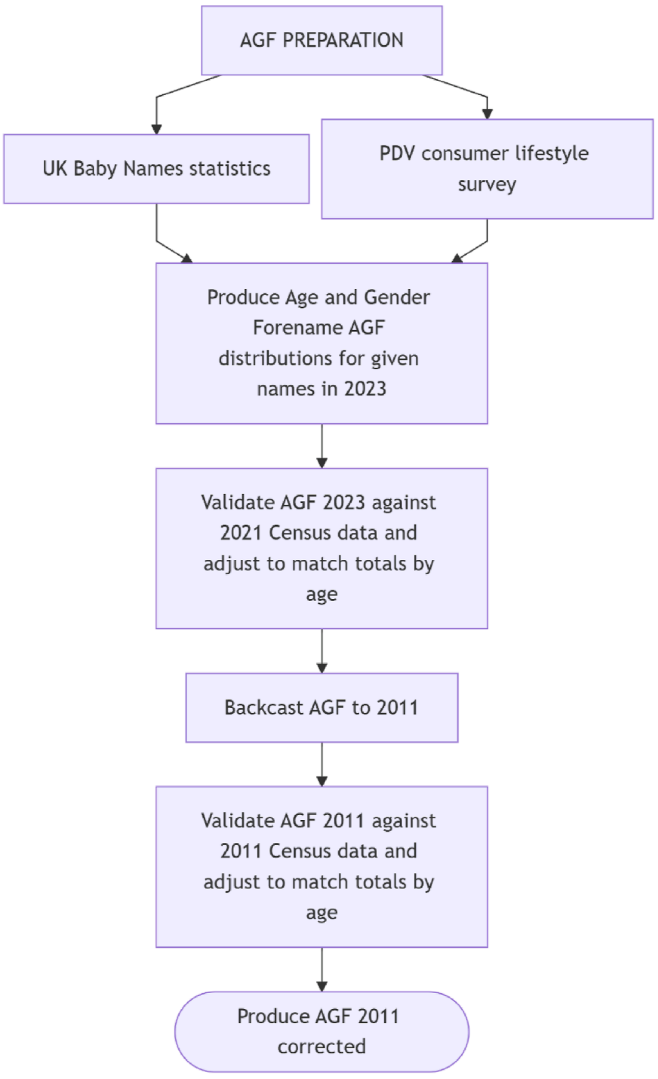


Fig. 1. Flowchart for the creation of the Age and Gender by Forename (AGF) distributions.

4. Refining 2011 LCR age and gender estimates

The reweighted 2011 AGFs for every given name were used to provide a range of provisional estimates of the ages of household members recorded in the 2011 LCR. In applying the 2011 AGF model, we were mindful that some forenames regain their popularity enjoyed in past generations, manifest in bi- or multi-modal AGF distributions. In developing a parsimonious implementation of the AGF model, we began as agnostic as to whether mean, median or modal values should be assigned to a name bearer, where there was no available ancillary information to refine estimates. Where such information was available, specifically from cohabitation with other adults, we were predisposed to use the median of candidate ages. We used the following steps to assign and validate ages for individuals in a range of household circumstances:

1. For households comprising a lone adult, the mean, median and modal ages were assigned from the forename age probability distribution.
2. For individuals living in multi-adult households, we estimated individual ages using different measures of central tendency and household heuristics developed from previous research, as set out in Table 1.

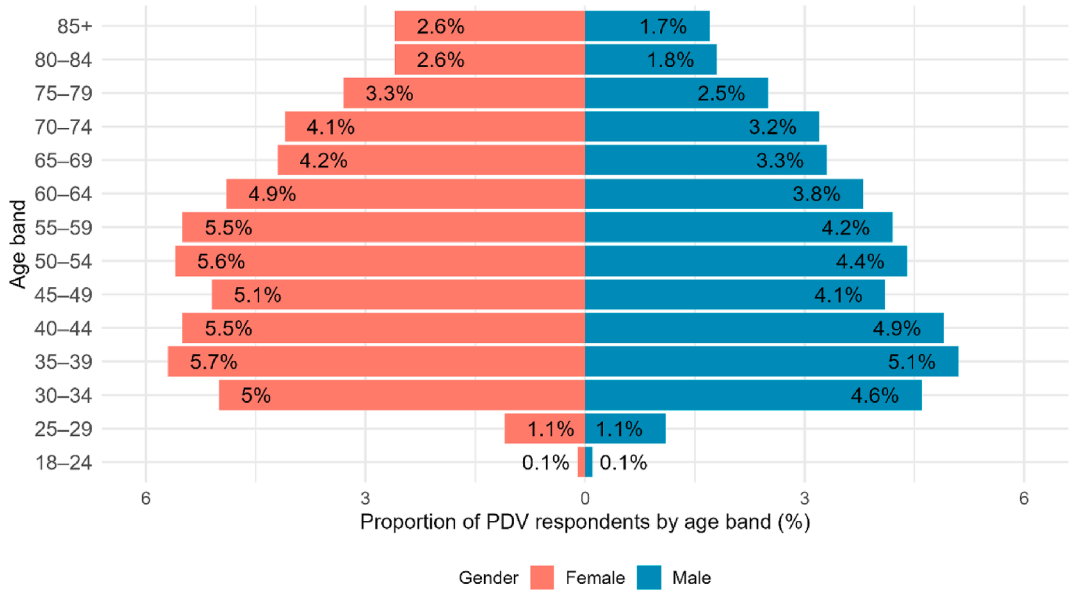


Fig. 2. Population age pyramid of respondents to the PDV surveys.

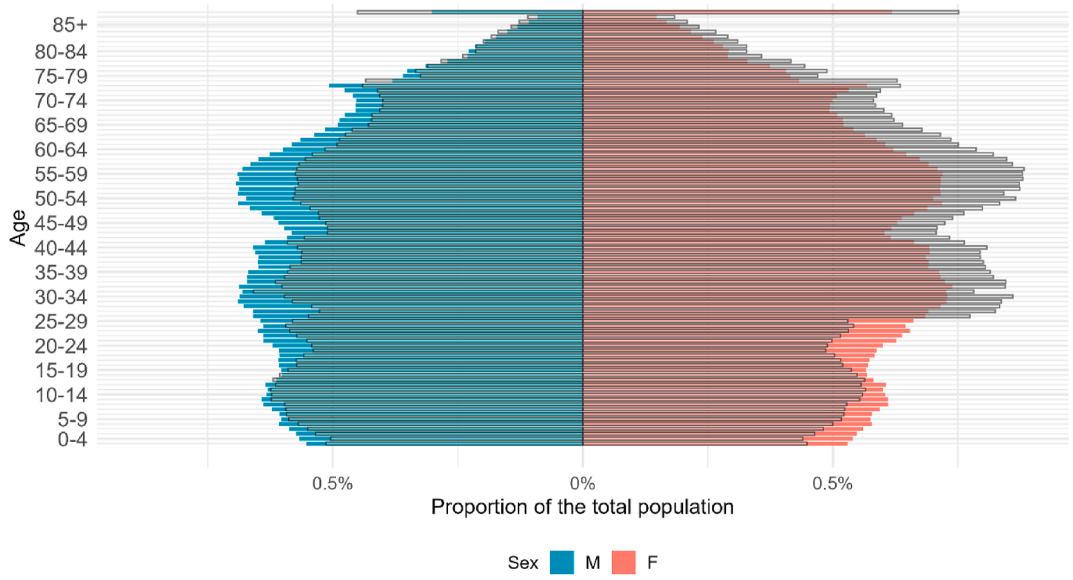


Fig. 3. Age pyramid for the 2023 Age and Gender by Forename (AGF) distribution (shown using black outlines) and harmonised 2021/2 Census estimates (solid colouring).

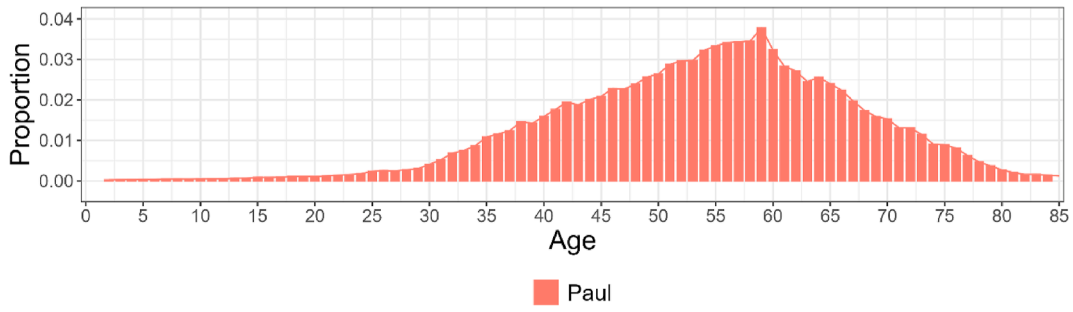


Fig. 4. Mortality adjusted probability density for the name 'Paul' in 2023, derived from the AGF (source: PDV Ltd combined with ONS Baby Names and ONS Life Tables).

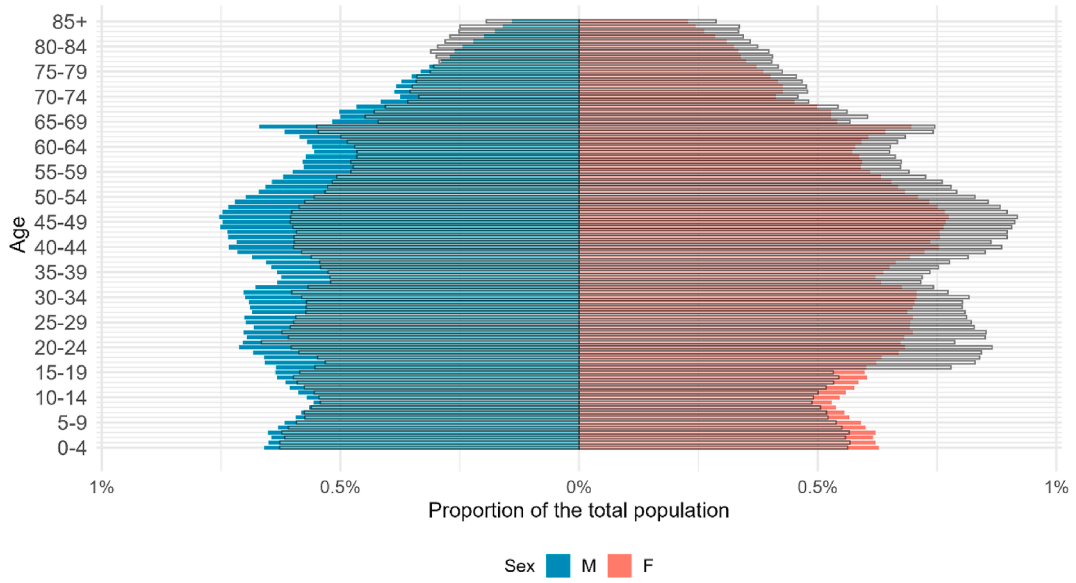


Fig. 5. Age pyramid for the 2011 Age and Gender by Forename (AGF) distribution (shown using black outlines) and 2011 Census estimates (solid colouring). Differences are used to devise age-specific weights to align the AGF with the Censuses.

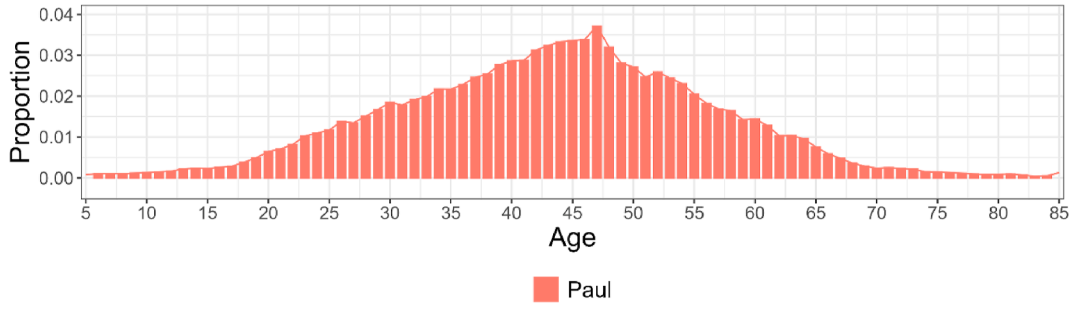


Fig. 6. Probability density distribution for the name 'Paul' in 2011, adjusted for age-specific mortality rates.

3. All 2011 LCR age estimates were then aggregated to LSOA level and compared with 2011 Census estimates.
4. For unique individuals identified in both the 2011 LCR and the PDV survey, estimated and observed ages were compared.

The heuristics set out in Table 1 are developed around what is known about names, ages and household structure in the UK. Specifically: (a) mean age differences between partners upon first marriage in 2011 was 2 years (ONS, 2013) and most differences were less than 10 years; (b)

Table 1
Single and multi-person household age calibration heuristics.

Number of individuals in household	Gender (M or F) of individuals in household	Surnames within household	Modal age difference/range between household members	Age adjustment
1	M or F	Any	Not applicable	Age candidate (mean, median, mode) from AGF 2011
2	M and F	Same surname	Less than 10 years	Household distributions with less than 2000 bearers of the two names Age female = mean of the inverse variance weighted means of the individual name distributions Household distributions with more than 2000 combined counts Age female = mode of the combined distribution In both cases Age male = Age female + 2
2	M and F	Any	More than 10 years	Retain individual (mean, median, mode) ages.
2	MM or FF	Any	Not applicable	Retain individual (mean, median, mode) ages.
3 or more	Any	2+ share same surname	Same surname difference less than 16 years	Take the two highest modes and use their combined distribution (?) and retain mean, median and mode for the remaining individuals in the household
3 or more	Any	2+ share same surname	Same surname difference more than 16 years	Take the second and third highest modes in the combined distribution and use age candidate of individual distribution for the remaining individuals in the household.
3 or more	Any	Different surnames		Age candidate (mean, median, mode) from AGF 2011

Table 2
Household composition counts and proportions of the number of people per household in the LCR and in the corresponding ONS table for 2011.

Number of adult individuals in household	LCR count	ONS count	LCR %	ONS %
1	8,687,009	7,875,613	40.067%	36.324%
2	8,629,116	8,794,219	39.800%	40.561%
3	2,825,772	3,984,040	13.033%	18.375%
4	1,090,855	3,221,192	5.031%	14.857%
5	287,371	872,978	1.325%	4.026%
6	80,571	162,456	0.372%	0.749%
7	28,263	21,614	0.130%	0.100%
8+	52,482	18,747	0.242%	0.086%

modal ages are most appropriate for characterising common names (defined through our analysis as having 1000+ bearers in the PDV survey) because some are observed to be multi-modal, while means provide better measures of rarer names that may have discontinuous distributions; and (c) 16 years is commonly taken as the minimum interval separating successive generations. Joint age distributions of co-habiting household members were obtained by combining the age distributions of the forenames associated with all permutations of different pairs of household members (Hancock et al., 2003). The most common operations were applied to two-person households in which members shared a common surname, with the aim of differentiating between households from the same generation (couples, siblings and

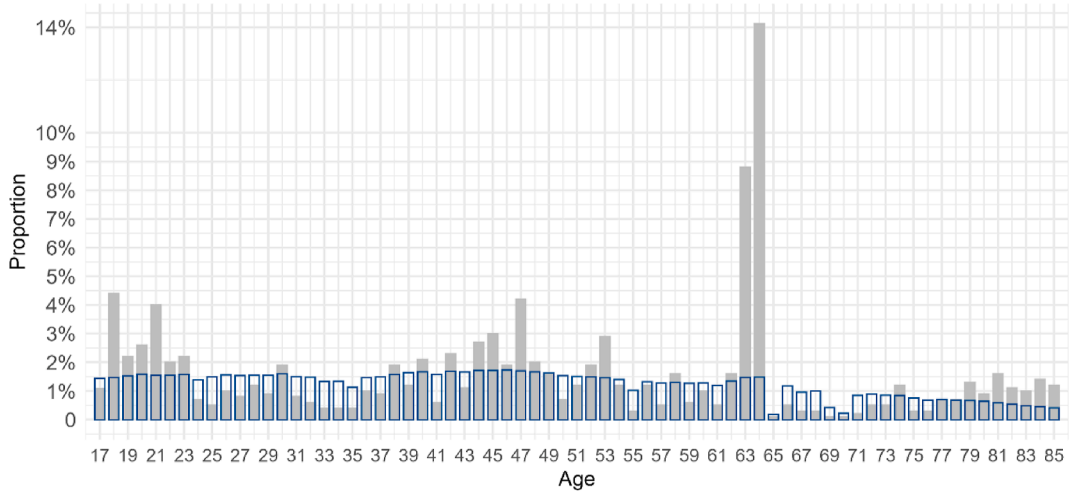


Fig. 7. Ages of individuals resident in single adult households, estimated by applying the mode of the AGF distribution to the 2011 LCR (shaded grey) and the 2011 Census (unshaded), expressed as proportions of the total population.

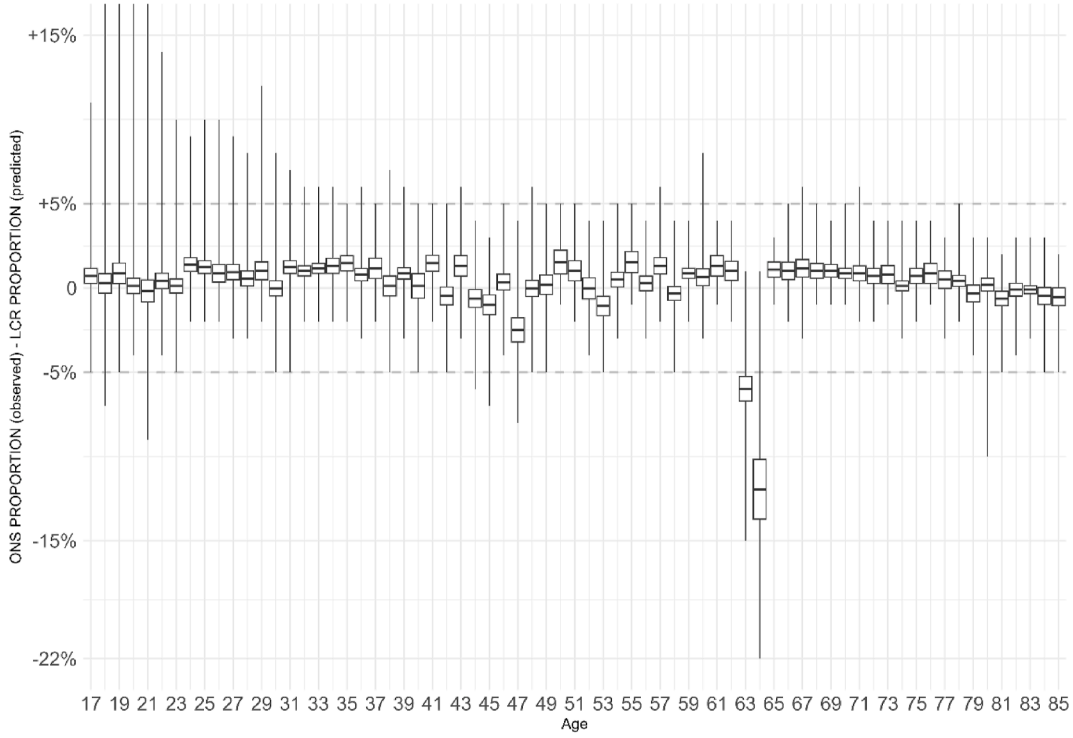


Fig. 8. Annual age band proportion residuals for one-person households in LCR 2011 using modal ages.

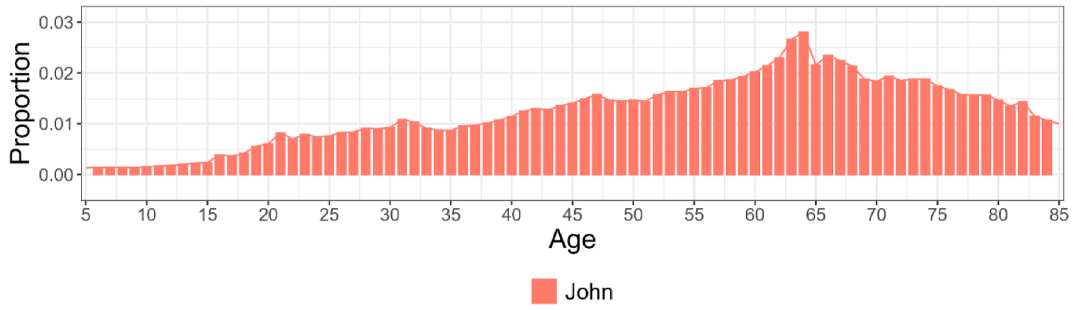


Fig. 9. Probability density distribution for the name 'John' in 2011.

informal households) and single parents living with non-dependent next-generation adults.

Households were defined simply as the total number of adults resident at an address in the 2011 LCR. This fulfils only the first elements of the ONS household definition of 'one person living alone, or a group of people (not necessarily related) living at the same address **who share cooking facilities and share a living room, sitting room, or dining area**' (emphasis added). The 2011 LCR significantly under-enumerates the number of adults living in households that include three, four or five adults (Table 2).

4.1. Model step 1: individual models

First, all individuals in single adult households were assigned the mean, median and modal adult age associated with their forenames from the 2011 AGF. These statistics were assigned for each of the 19,237 unique forenames that have 100+ bearers in the PDV data, accounting for 99+% of all records in the 2011 LCR. Fig. 7 presents the distribution of modal AGF estimates which, in common with the other measures of central tendency shows bunching of AGF estimates in early adulthood and late old age. The coincidence of the modal age of some very common names (e.g., John or Patricia) at ages 63 and 64 leads to heavy over-estimation of these age cohorts, contrasting with persistent, albeit much less pronounced, under-estimation throughout much of the distribution spanning ages 24–78.

Fig. 8 presents the distribution of residuals for each annual age band at the LSOA level, calculated as the difference between the proportion of named individuals in each predicted age class in the LCR and the corresponding 2011 Census estimates. The first five years of adulthood aside, most of the residuals are close to zero, falling within±5% of the Census estimate (shown by the grey dotted lines), with the exception of ages 63 and 64, where extremely low Census values manifest the distorting effect of several very common names with wide age distributions but shared modal values. Fig. 9 illustrates how the name 'John' contributes to this effect, remaining common across the 2011 AGF, with a small peak at age 64: the popularity of the name peaked in the 1950s and has subsequently declined incrementally. In the absence of household or other adjustments, this means that the modal value has a strong influence on the overall modelled age distribution.

Fig. 10 further illustrates the distorting effects upon 2011 LCR AGF estimates of bearers of forenames with modal ages of 18, 63, and 64. It is evident that modal ages of common forenames such as John, Robert, Peter, Mary, Patricia, Paul or Andrew provide limited information content and have wide error margins. In households comprising multiple adults, these effects can be mitigated by adjusting modal values utilising information from other household members.

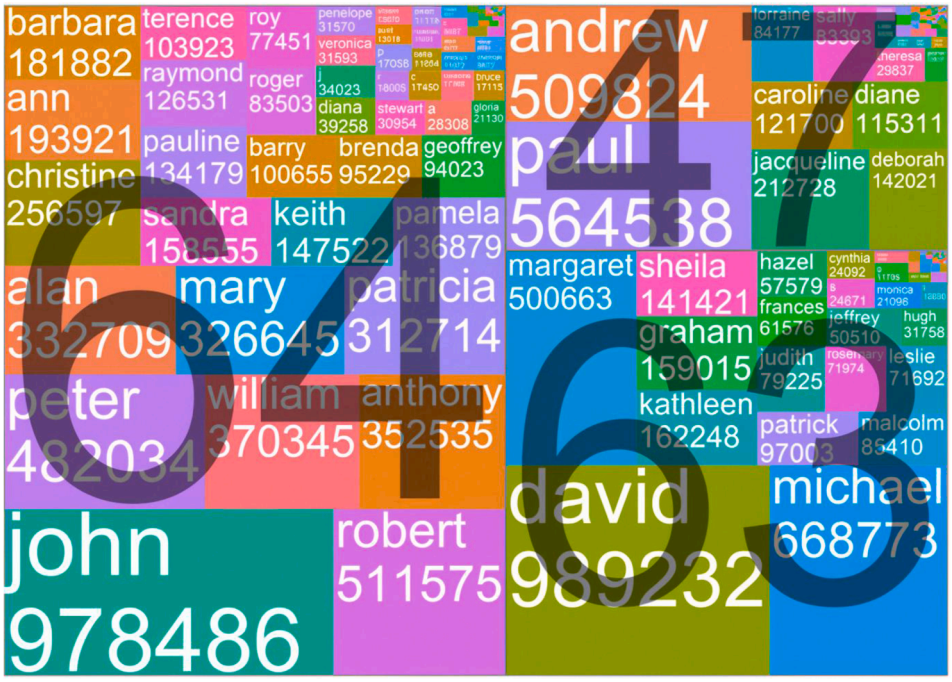


Fig. 10. Treemap chart of 2011 PDV forenames with modal ages 63, 64, and 47.

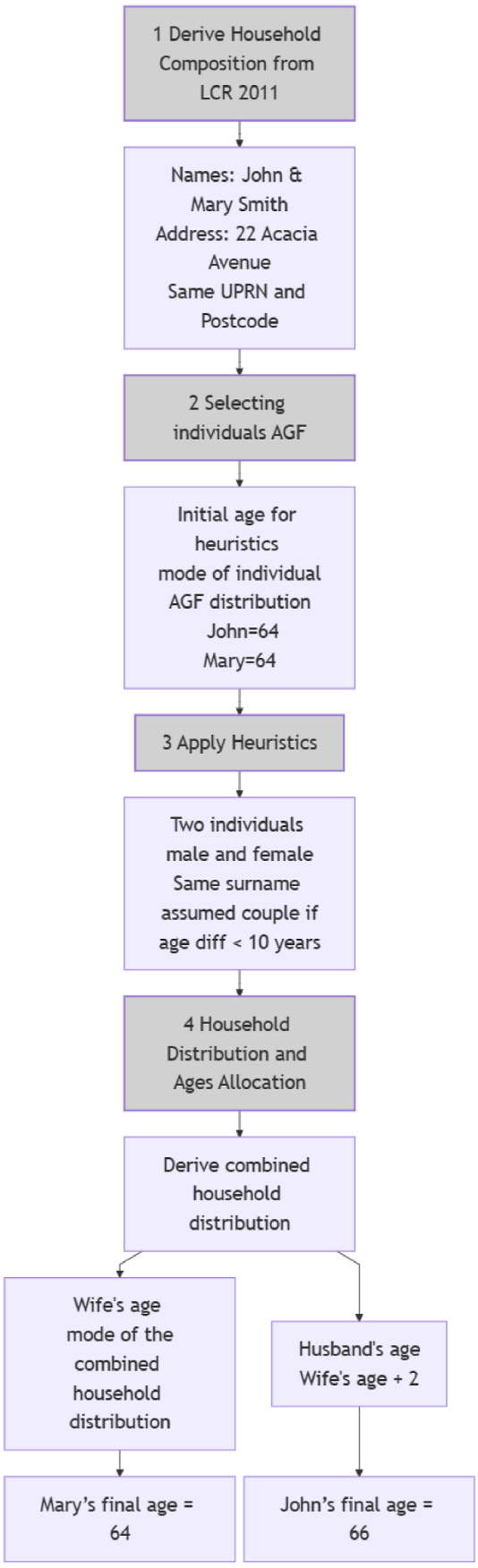


Fig. 11. 2011 Age estimation for John and Mary Smith living together.

4.2. Model step 2: household model

The household model was used to adjust the estimated ages of cohabiting individuals to improve consistency and precision of within-household age estimates. Adjusted age estimates were obtained by applying the heuristics shown in Table 1, as appropriate, to multi-person households, and comparing multiple AGF summations where necessary. The heuristics are based on the number of persons in each household with a focus on differentiating between co-habiting couples (with or without dependent children), intergenerational families and informal multi-adult households as formulated by Ni Bhrolcháin (2005), Hancock et al. (2003) and our own experimentation. Fig. 10 illustrates the approach for a household comprising John and Mary Smith (Table 1, case 2). Two-adult households account for approximately 40% of the records in the LCR (Table 2), while John and Mary are two very common names in the UK (Fig. 10). Fig. 11 confirms that John and Mary are more likely to be partners than an intergenerational family. The heuristic for this case in Table 1 suggests calculating the difference between the ages of the individuals derived from the AGF distributions: the difference is less than 10 years and so we considered the two individuals to be a couple. The names John and Mary together have more than 2000 observations and so we calculated the inverse variance weighted mean of their combined distributions. The adjusted age for the female, Mary, was the inverse variance weighted mean of her age. The age of the husband was then calculated as the age of the wife plus two years to reflect the typical difference in age upon marriage (Hancock et al., 2003).

For cases where the number of observations making up the 2011 AGF were less than 1,000, such as for Zafar and Shreya (Fig. 12), the age of the female was calculated as the mode of the household distribution, 18, and the age of the husband is therefore 20.

5. Results and validation

We assessed whether the addition of heuristics would improve the age estimation model and then chose the available options that performed best. Analysis of heuristics is computationally intensive and, given the requirement to work with personal data in a secure trusted research environment, it was not possible to undertake evaluation using the entire population data. Evaluation of the available heuristics was therefore performed across three contrasting boroughs: Camden, Solihull and Knowsley. Table 3 summarises their principal socio-economic and demographic characteristics according to the 2011 Census and Output Area Classifications (OAC: Gale et al., 2016). Solihull was selected because of its predominantly suburban character, with a significant proportion of detached and semi-detached dwellings, and its relatively aged demographic profile. This selection reflects the borough's appeal as a residential area for retirees. Camden is a densely populated, ethnically diverse inner-city area with a younger, more heterogeneous population, including a significant proportion of students, young professionals, and immigrants. Knowsley was selected for its predominantly working-class suburban profile, with a younger population and a significant proportion of social housing. The Borough has low educational attainment and high unemployment compared to the UK national average. In contrast to Solihull's older demographic, Knowsley has more families with children and a less ethnically diverse population, with over 95% identifying as White British.

The validation process encompassed three dimensions: LSOA level, aggregated age level, and individual level, with the goal of assessing the best measure of central tendency from the AGF and the value of using the household heuristics. The final step of the validation process involved matching LCR individuals in the three local authorities to the original PDV survey.

First, we compared the aggregated LSOA estimates to the 2011 Census. The Census tables enable calculation of proportions of the LSOA falling into every adult single year of age. These were compared with LSOA aggregations of age estimates calculated using the AGF means,

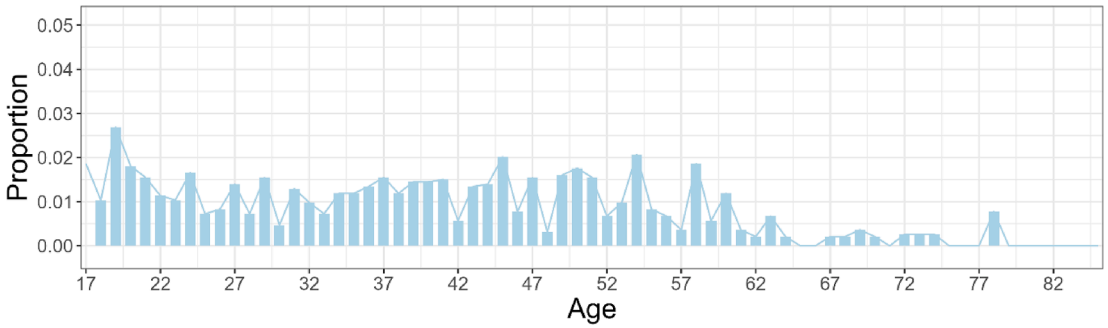


Fig. 12. Household distribution for the names Zafar and Shreya.

Table 3
Socio- and geo-demographic characteristics of Camden, Solihull and Knowsley in 2011 (source: ONS Census statistics).

	Camden	Solihull	Knowsley
Total Population	220338	206674	145893
Male (%)	48.9	48.6	48.3
Female (%)	51.1	51.4	51.7
Aged 0–15 (%)	16.1	18	19.7
Aged 16–64 (%)	73.1	61.6	64.4
Aged 65+ (%)	10.8	19.2	15.9
White (%)	66.3	89.1	97.2
Other Ethnicities (%)	33.7	10.9	4.7
OAC Supergroups	Cosmopolitan, Multicultural Metropolitans	Suburbanites, Rural Residents	Hard-Pressed Living, Suburbanites
OAC Supergroups characteristics	Younger populations, high rental density, diverse student and professional communities	Older populations, homeownership, affluent suburban and rural areas	Higher unemployment, younger population, social housing, family-oriented suburban areas

medians and modes of the forenames of all adults present in the LCR. Corresponding calculations were also made using the heuristics for multi-adult households as set out in Table 1. In each instance, model performance was assessed by calculating: the root mean squared error (RMSE) differences in proportions; R^2 statistics obtained from regressing predicted LSOA mean age against observed mean age; and counts of the total number of predicted age bands found (with higher counts seen as indicating less distorting concentration of results on the modal ages of bearers of common names). The data in Table 4 confirm that the introduction of heuristics improved over-all model performance for every measure of central tendency. The RMSE is smallest, while R^2 and

Table 4
Model performance for age estimation models with and without the use of heuristics for Camden, Solihull and Knowsley.

AGE CANDIDATE	RMSE PROP AGE TOTALS	R-SQUARED AGE TOTALS	AGE CLASSES IDENTIFIED
Mean without heuristics	0.0109	0.3037	58
Mean with heuristics	0.0091	0.3806	69
Median without heuristics	0.0098	0.3326	67
Median with heuristics	0.0086	0.3661	69
Mode without heuristics	0.0280	0.0363	69
Mode with heuristic	0.0208	0.0541	69

Table 5
Model performance metrics for all age estimation models with the use of heuristics: Camden, Solihull and Knowsley.

	MEAN LSOA RMSE	MAE PDV INDIVIDUAL	CORRECT AGE PROP (%)	AGE PLUS OR MINUS 1 (%)
Mean with heuristics	0.0124	10.422	3.40%	10.30%
Median with heuristics	0.0116	10.4686	3.80%	10.85%
Mode with heuristic	0.0227	11.8685	3.78%	10.77%

age class count statistics suggest that use of the AGF median provides the closest fit with the Census.

Second, we calculated similar RMSE differences between the proportions falling into every adult age band in the study areas (Table 5, column 1). Here again the use of AGF medians with heuristics enabled the closest match with the Census LSOA tables. Third, the individual-level validation yielded mixed results, which had been anticipated given that the heuristics were applicable only to a subset of the population.

The use of medians with heuristics performed consistently well across all of the metrics, delivering the best values for R^2, RMSE and the proportion of PDV-matched individuals whose ages were precisely estimated. Fig. 13 compares the recorded ages of PDV survey respondents with the mean of their estimated ages using the AGF model. AGF estimates gravitate towards middle age and tend to be most adrift of recorded ages in the youngest and oldest cohorts, consistent with Fig. 8.

6. Discussion and further work

The opening section of this paper argued that new sources of smart data enabled integration of richer and more frequently updated data into information geographies, but that this required data augmentation, linage and triangulation to realise these benefits. It also argued that these processes were best accomplished at the level of the human individual to avoid issues of ambiguity and ecological fallacy in aggregated data. Our empirical analysis has demonstrated how individual level smart data can be augmented, and predictive errors managed, using probabilistic distributions of ages developed from ancillary sources and validated using aggregate social survey (census) data. Retention of the human individual as the unit of modelling and analysis in these endeavours renders the results flexible in further linkage to other smart, survey or administrative sources. From this perspective, smart data enable enrichment through linkage to other information geographies, exploiting point geographical referencing to bring together diverse smart data from different data domains. Estimation of individual age and gender estimates in ISO27001 environments also means that the approach preserves individual privacy, since stand-alone or lined estimates are only exported in aggregated form, subject to similar

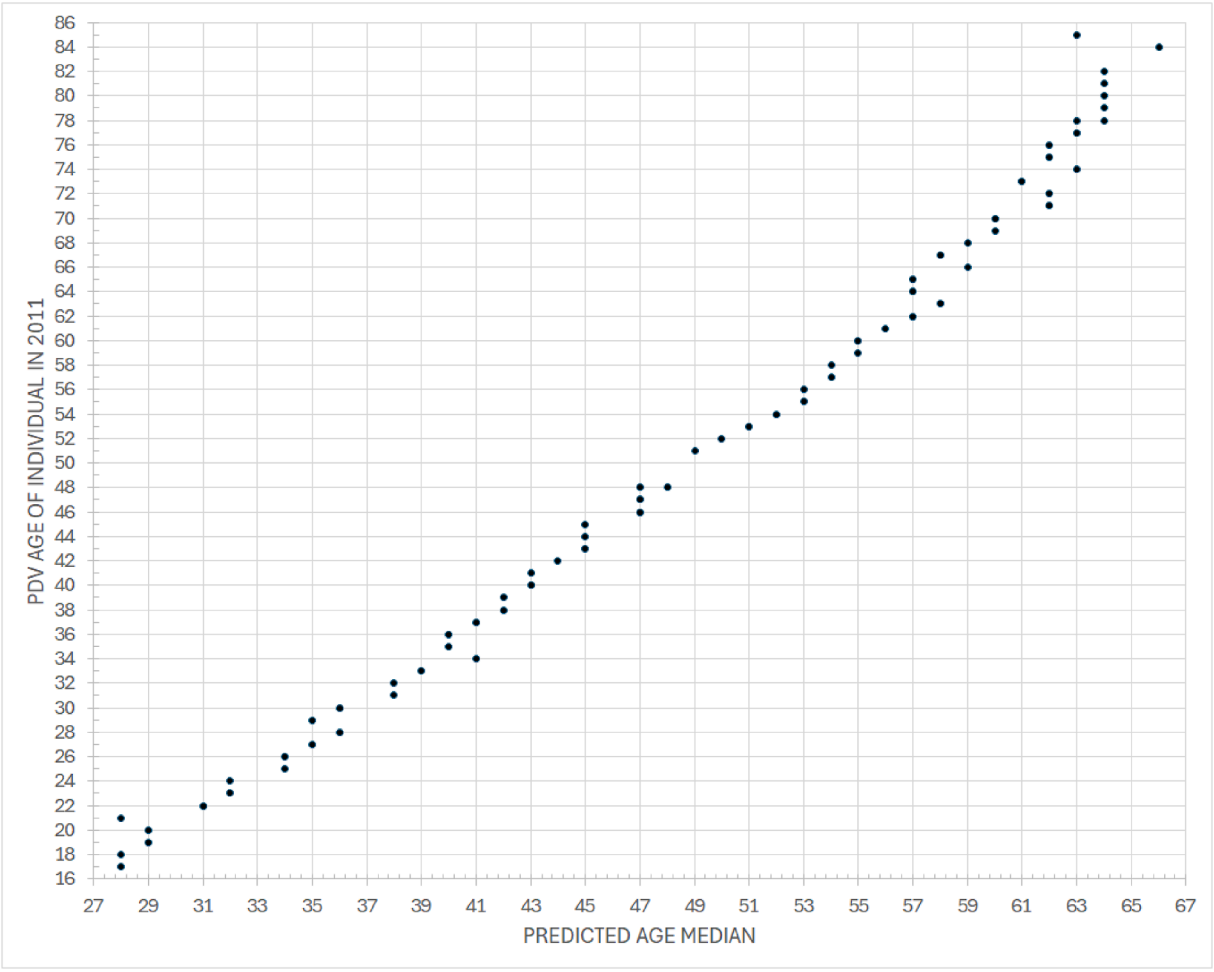


Fig. 13. Median values for individuals' estimated ages compared to values recorded in the Lifestyles Survey using median estimates and heuristics: Camden, Solihull and Knowsley.

thresholds to those used by the U Office for National Statistics.

This said, estimating an individual's age solely based on their forename presents significant challenges, particularly in the absence of supplementary information. In this study, we developed and applied heuristic methods to adjust age estimates by considering the characteristics of the household in which the individual was resident. We conceived this as a general-purpose problem framed only by numbers of adults in each household, and so refrained from using any explicit spatial model, such as stratifying the AGF distribution by geography – at any scale ranging from the regional to the very local. Our findings indicate that the use of heuristics significantly enhances the accuracy of age and sex estimations, outperforming use of measures of central tendency alone. The goals of reproducing neighbourhood age structures as measured by censuses and of predicting individuals' ages to within one year are very ambitious: our results are encouraging, but further research is required.

A critical limitation of the data and the model is posed by single adult households. These are over-represented in the LCR, and we have not devised heuristics to adjust our estimates beyond forename-specific measures of central tendency. Future research might develop and use contextual information to supplement household composition. This might include, for example, construction type of residence, neighbourhood characteristics such as residential density, duration of residence, or other linkable data. Most fundamentally, perhaps, might be to use precise georeferencing and the geodemographics adage that 'birds of a feather flock together' (Harris et al., 2005) to validate age estimates with respect to adjacent or neighbouring properties.

Although the method reported here drives the focus of estimation to the ultimate level of the individual, we have not attempted to accommodate the geographic context in which any individual lives. This may be important in potential applications, since there are regional as well as social variations in naming practices, present day and historic. We have not attempted to accommodate these here, since they presume data about and understanding of both regional naming practices and the cumulative effects of subsequent residential mobility patterns. Related to this, regional and social variations in the UK-wide mortality statistics that we use to hindcast the 2011 age distribution. In terms of the motivations for this research, these considerations are not of primary concern in the development of name-specific age distributions, but present a focus for future research to create an individual level foundation model for forecasting social, economic and demographic change in the UK. Detailed understanding and accommodation of the socio-economic biases in the PDV data will be integral to this task. A further extension might be to hindcast name- and gender-specific age distributions beyond the current generation, perhaps using historical telephone directory or census data (Tanu et al., 2024; Lan and Longley, 2023).

These approaches introduce considerable additional complexity and must be pursued with caution. Refining both individual and joint household models and carefully weighting the predictions for consistency with statistical sources can enable highly disaggregate local models that can be used to predict changing neighbourhood circumstances over the short, medium and longer terms. The associated processes of data triangulation and augmentation should be carefully documented to make clear the mix of general purpose versus contextual

modelling that has been undertaken. The continued exploration of name-based predictions offers significant potential, particularly when addressing the challenges posed by uneven data distributions and varying household compositions. Attribution of age estimates to precisely georeferenced addresses enables non-disclosive aggregation of results to any convenient administrative or grid geography (see, for example, Lloyd, 2015; Martin et al., 2013). Smart data today offer great depth of behavioural insight through the provision of detailed transactions or interactions with smart devices, but little or no information as to how the data relate to any known population. Inference of generalisable demographic characteristics such as age and gender enable triangulation of such data sources (e.g. see Grow et al., 2022). Triangulation of aggregated estimates to grids or administrative geographies also enables short-term updating and monitoring of change dynamics where conventional statistical or administrative sources are rarely updated and may be required for any of a range of aggregations of individuals.

CRediT authorship contribution statement

Maurizio Gibin: Writing – original draft, Methodology, Formal analysis. **Justin van Dijk:** Writing – review & editing, Methodology, Conceptualization. **Zi Ye:** Writing – review & editing. **Paul Longley:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Funding

We acknowledge funding from the Economic and Social Research Council (ESRC) via the Geographic Data Service (GeoDS) grant, reference ES/Z504464/1 (Gibin, Ye, Longley).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paul Longley reports financial support was provided by Economic and Social Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Gale, C.G., Singleton, A.D., Bates, A.G., Longley, P.A., 2016. Creating the 2011 area classification for output areas (2011 OAC). J. Spat. Inf. Sci. 1–27. https://doi.org/10.5311/JOSIS.2016.12.232.

Goodchild, M.F., 2022. Elements of an infrastructure for big urban data. Urban Informatics 1, 3. https://link.springer.com/article/10.1007/s44212-022-00001-5.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R.D., Ventura, I., Weber, I., 2022. Is Facebook's advertising data accurate enough for use in social science research? insights from a cross-national online survey. J. Roy. Stat. Soc. 185 (S2), S343–S363. https://doi.org/10.1111/rssa.12948.

Hancock, R., Stuchbury, R., Tomassini, C., 2003. Changes in the distribution of marital age differences in England and Wales, 1963 to 1998. Popul. Trends 19–25.

Harris, R., Sleight, P., Webber, R., 2005. Geodemographics, GIS and Neighbourhood Targeting. Wiley, Chichester.

Kandt, J., Van Dijk, J., Longley, P.A., 2020. Family name origins and intergenerational demographic change in great Britain. Ann. Assoc. Am. Geogr. 110 (6), 1726–1742. https://doi.org/10.1080/24694452.2020.1717328.

Kandt, J., Longley, P.A., 2018. Ethnicity estimation using family naming practices. PLoS One 13 (8). Article e0201774. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201774.

Lan, T., Kandt, J., Longley, P.A., 2020. Geographic scales of residential segregation in English cities. Urban Geogr. 41 (1), 103–123. https://doi.org/10.1080/02723638.2019.1645554.

Lan, T., Longley, P.A., 2023. An individual level method for improved estimation of ethnic characteristics. Int. Reg. Sci. Rev. 46 (3), 328–353. https://doi.org/10.1177/01600176221116568.

Lansley, G., Li, W., Longley, P.A., 2019. Creating a linked consumer register for granular demographic analysis. J. R. Stat. Soc. Ser. A Stat. Soc. 182, 1587–1605. https://doi.org/10.1111/rssa.12476.

Lansley, G., Longley, P.A., 2016. Deriving age and gender from forenames for consumer analytics. J. Retail. Consum. Serv. 30, 271–278. https://doi.org/10.1016/j.jretconser.2016.02.007.

Lloyd, C.D., 2015. Local cost surface models of distance decay for the analysis of gridded population data. J. Roy. Stat. Soc. 178 (1), 125–146. https://doi.org/10.1111/rssa.12047.

Longley, P.A., Chen, M., 2025. Smart data, information geographies and intelligent data services. Inf. Geogr. 1 (1), 100013. https://doi.org/10.1016/j.infgeo.2025.100013.

Longley, P.A., Singleton, A.D., Cheshire, J.A., 2024. 'Research ready' geographically enabled smart data. Spatial Sci. 1–7. https://doi.org/10.1080/19475683.2024.2353035.

Longley, P., Singleton, A., Cheshire, J., 2018. Consumer Data Research. UCL Press. https://doi.org/10.14324/111.9781787353886.

Martin, D., Cockings, S., Harfoot, A., 2013. Development of a geographical framework for census workplace data. J. Roy. Stat. Soc. 176 (2), 585–602. https://doi.org/10.1111/j.1467-985X.2012.01054.x.

McGrath-Lone, L., Jay, M.A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wijlaars, L., Gilbert, R., 2022. What makes administrative data research-ready? A systematic review and thematic analysis of published literature. Int. J. Popul. Data Sci. 7 (1). https://doi.org/10.23889/ijpds.v7i1.1718.

Ni Bhrolcháin, M., 2005. The age difference at marriage in England and Wales: a century of patterns and trends. Popul. Trends 7–14.

Office for National Statistics, 2013. Marriages in England and Wales (Provisional): 2011. ONS Statistical Bulletin. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/marriagecohabitationandcivilpartnerships/bulletins/marriagesinenglandandwalesprovisional/2013-06-26.

Office for National Statistics, 2024. National Life Tables: United Kingdom, 1980–1982 to 2020–2022. ONS Statistical Bulletin. https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesenglandreferencetables/current/previous/v8/nlte198020203.xlsx.

Openshaw, S., 1984. The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography 38. Geobooks, Norwich, UK.

Tanu, N., Gibin, M., Hu, D., Longley, P.A., 2024. A century of telephony: digital capture of British telephone directories, 1880-1984. Spatial Sci. 30 (2), 167–180. https://doi.org/10.1080/19475683.2024.2331509.

Van Dijk, J.T., Lansley, G., Longley, P.A., 2021. Using linked consumer registers to estimate residential moves in the United Kingdom. J. R. Stat. Soc. Ser. A Stat. Soc. 184, 1452–1474. https://doi.org/10.1111/rssa.12713.