Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–23
https://doi.org/10.1093/jrsssc/qlad095

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

Combining individual- and population-level data to develop a Bayesian parity-specific fertility projection model

Joanne Ellison[1], Ann Berrington[1], Erengul Dodd[2] and Jonathan J. Forster[3]

[1]Department of Social Statistics and Demography, University of Southampton, Southampton, UK
[2]School of Mathematical Sciences, University of Southampton, Southampton, UK
[3]Department of Statistics, University of Warwick, Coventry, UK

Address for correspondence: Joanne Ellison, Department of Social Statistics and Demography, University of Southampton, Highfield, Southampton SO17 1BJ, UK. Email: J.V.Ellison@soton.ac.uk

Abstract

Fertility projections are vital to anticipate demand for maternity and childcare services, among other uses. Models typically use aggregate population-level data alone, ignoring the richness of individual-level data. We hence develop a Bayesian parity-specific projection model combining such data sources. We apply our method to England and Wales, using individual-level data from *Understanding Society*. Fitting generalised additive models gives smooth projections across age, cohort, and time since last birth. We also incorporate prior beliefs about the relative importance of the data sources. Our approach generates plausible forecasts by individual-level variables including educational qualification, despite their absence in the population-level data.

Keywords: Bayesian methods, combining data sources, fertility forecasting, generalised additive models, parity

1 Introduction

Fertility projections carry great importance on their own, for example in planning maternity services and anticipating demand for school places. As a component of overall demographic change, they also have a highly significant influence on projected population sizes, age structures, and the associated uncertainty (United Nations Development Programme, Department of Economic and Social Affairs, Population Division, 2019a, 2019b). Despite the large and diverse literature concerning fertility forecasting models (see Bohk-Ewald et al., 2018; Booth, 2006 for reviews), these typically use aggregate population-level data alone, e.g. from vital registration. However, the intrinsic dependence of fertility on human decisions means that childbearing behaviour is influenced by a variety of factors acting at different levels, such as the individual or country level, with individual-level determinants including partnership status, income, employment, and upbringing (Balbo et al., 2013). Furthermore, the precise impacts of these factors differ by parity (Fiori et al., 2014), which is defined as the number of previous live-born children. Fertility projections by individual-level variables including education (e.g. Abel et al., 2016; Vollset et al., 2020) and ethnicity (e.g. Norman et al., 2014) have been produced, but these result from a specific desire to disaggregate by these covariates. In this paper, we propose a general methodology to incorporate individual-level variables into a parity-specific fertility projection model, determining the exact variables to be used within the modelling process.

Received: January 3, 2023. **Revised:** June 28, 2023. **Accepted:** August 31, 2023

An established body of literature is dedicated to modelling fertility at the individual level and exploring the impact of background factors such as those listed above. Surveys are typically used as they often provide detailed birth histories and individual-level background covariates. It is standard practice for such analyses to take parity into account, consistent with evidence that the factors influencing the transition to motherhood differ greatly from those affecting the transition to higher-order births (Fiori et al., 2014). The explanatory focus of much of the existing literature, together with the likelihood that the individual-level data will suffer from the issues of inaccuracy, bias and non-response, mean that predictive extensions remain unexplored. Conversely, published birth registration data has little detail apart from year of birth, mother's age and possibly parity; despite this, it is collected for all births and so avoids such sampling-related problems. In this way, the two levels of data complement each other, with a shortcoming of one being a strength of the other. The intuitive next step is to develop a fertility forecasting model that combines them in some way; to the best of our knowledge, a general approach to achieve this does not exist in the literature.

The proposal to combine individual- and population-level data in itself is not new. Early work views the population-level data as exact, and thus able to increase the accuracy and precision of parameter estimates from models fitted to individual-level data. Imbens and Lancaster (1994) use generalised method of moments estimation to specify moment conditions that force the population-level constraints to be satisfied; Handcock et al. (2000) present a constrained maximum-likelihood approach to the same end. Handcock et al. (2005) extend their earlier work by constraining to subgroup data as opposed to simply the general fertility rate; this improves the efficiency of coefficient parameter estimates as well as the intercept. Rendall et al. (2008) use the same method, finding that pooling surveys can even improve the efficiency of indirectly constrained parameter estimates. The proposed empirical-likelihood-based method of Chaudhuri et al. (2008) only requires the satisfaction of linear constraints that do not depend explicitly on the model parameters, thus providing greater computational efficiency than solving non-linear constraints in the approach of Handcock and colleagues.

Bayesian approaches allow the integration of prior knowledge about the constraints and can therefore account for bias in the population-level data. In the context of fertility, Rendall et al. (2009) identify upward bias in the population-level age-specific fertility rate (ASFR) estimates for Hispanic women in the USA, partly due to census undercounts of the risk population. Their Bayesian constrained maximum-likelihood model fitted to survey data specifies various elicited priors for the population-level constraints, allowing for different amounts of correction and uncertainty. Contrastingly, Zhang and Bryant (2019) identify downward bias in the population-level province-specific ASFR estimates for Cambodian women due to census undercounts of births. Their hierarchical Bayesian model for the true counts incorporates measurement error models for the census and survey counts, enabling quantification of the undercoverage of the former and allowing the latter, assumed unbiased, to drive the model estimates. Such studies demonstrate how different levels of data and their associated uncertainties can be combined, and known biases can be incorporated, coherently within a Bayesian framework. They are also consistent with the increasing popularity of Bayesian methods in demographic forecasting (e.g. see Bijak & Bryant, 2016).

The advent of methods that can handle bias in the population-level data is particularly relevant in the case of our population-level data source, namely parity-specific fertility rate estimates for England and Wales from the Office for National Statistics (ONS, 2020). Before 2012, information about previous births was only requested at the registration of marital births (Smallwood, 2002). Furthermore, only children with the current husband and any previous husbands were considered (ONS, 2022b). Therefore, the registration birth order (when collected) differed from the true birth order, complicating the calculation of parity-specific fertility rates. To combat this problem, ONS used data from fertility histories collected within an annual sample survey, the General Household Survey, to adjust for the missing data (ONS, 2022b; Smallwood, 2002). In this way, there is inherent uncertainty and potential bias in the ONS (2020) population-level estimates, meaning that they cannot be viewed as the gold standard; this strengthens our motivation for a combined model.

Existing fertility projection models generally neglect individual-level influences, and despite parity-specific fertility data being collected by most countries (Jasilioniene et al., 2015), also ignore such information. This paper addresses these gaps in the literature by developing a Bayesian parity-specific fertility projection model that combines individual- and population-level data,

thereby exploiting the richness of this information. We illustrate our method in the context of England and Wales, but it can be applied to any country with the required data. Through fitting generalised additive models (GAMs) to the individual-level data, we estimate smooth effects of the covariates on the likelihood of a subsequent birth. We then determine suitable covariate models to marginalise over the variables unique to the individual-level data and thus integrate the population-level data. Our use of weights to balance the information in the data sources is a key methodological innovation of our approach. The paper proceeds as follows. In Section 2, we describe the data sources that we use to fit our model, which we specify in terms of the individual-level and combination stages in Sections 3 and 4, respectively. We present the results in Section 5, including an assessment of their sensitivity to the precise weight chosen, and provide a discussion in Section 6. The paper builds on work undertaken by Ellison (2021).

2 Data sources

Our individual-level data comes from Wave 1 of *Understanding Society*, i.e. the UK Household Longitudinal Study (UKHLS) (University of Essex Institute for Social and Economic Research, NatCen Social Research, Kantar Public, 2017), which collected fertility histories and additional information from 27,792 women between 2009 and 2011 (Institute for Social and Economic Research, 2022). Our sample consists of 18,218 of these women, born between 1945 and 1992, who provided complete and valid fertility histories and were living in England or Wales at the time of the interview. The sample selection process is described in online supplementary Appendix A. From each fertility history, we construct a series of records, one for each year of age the woman is observed at from 15 to 44. Each record has a corresponding binary response variable indicating whether the woman had a birth event. It additionally includes the values of the 'clock' variables (Raftery et al., 1996), i.e. age, period, cohort, time since last birth and parity, and any extra survey covariates. We have 357,287 records in total across the sample.

Our population-level data consists of parity-specific fertility rates estimated by ONS (2020). The dataset covers parities 0–4+ for the 1920–2003 birth cohorts from ages 14 to 45. The last year of observation is 2018 (compared to 2008 for our UKHLS sample), so for example the 2003 cohort is observed up to age 15. For consistency across datasets, we henceforth restrict our analyses to the 1945–2003 cohorts, from ages 15 to 44 and at parities 0, 1, 2, and 3+.[1] We compare the data sources, concentrating first on their observed aggregate and parity-specific rates (Figure 1). For a given age and cohort group, we compute the former by dividing the total number of births by the total number of women, and the latter for parity i by dividing the number of births of order $i+1$ by the number of women at parity i (taking $i+1$ to be 4+ when $i=3+$). For the UKHLS data, we weight the records using cross-sectional Wave 1 survey weights, which are specific to each individual and designed to improve the representativeness of analyses. We ignore parity–age–cohort combinations in the ONS data if the exposure estimate is low (negative, zero, or positive but exceeded by the birth count), which only occurs in the higher parities at the youngest ages.[2]

The most striking difference between the UKHLS rates and their ONS counterparts is the substantially increased smoothness across age and cohort in the latter, owing to the ONS exposures being on average 1,000 times larger than the corresponding UKHLS exposures. The significant noise present in the UKHLS rates, which obscures many of the shapes and trends, is mostly removed in the ONS rates. However, small exposures at the youngest ages for women who have already had a birth mean that the ONS rates remain erratic here, while the UKHLS rates are typically unobserved or based on even smaller exposures. In Section 3, we describe the first stage of our modelling process involving the UKHLS data only.

3 Modelling the individual-level data

3.1 Introduction

In this section, we determine appropriate models for the individual-level data, considering parities 0–3+ separately. We fit GAMs (Wood, 2017), a flexible class of models in which main effects and

[1] We combine parities 3 and above due to small cell counts at higher parities in the UKHLS data.

[2] In our subset of the ONS (2020) data there are only two instances of negative exposures, which are said to be anomalies occurring during data processing; there are 28 occasions where a positive exposure is exceeded by the birth count.

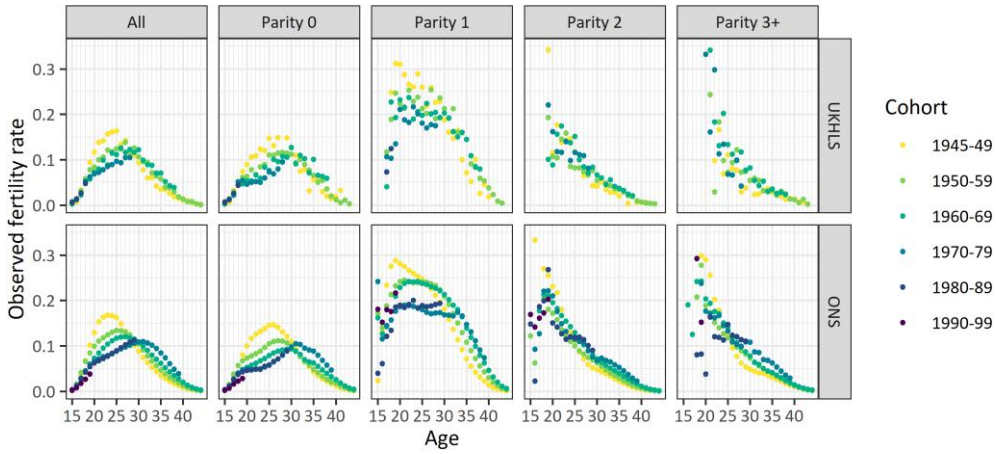

Figure 1. Observed fertility rates for England and Wales by parity, year of age, and grouped cohort year of birth, from the UKHLS and ONS data sources. UKHLS rates calculated using the records in our sample (weighted); ONS rates calculated using ONS (2020). Note that rates with small underlying counts have been omitted. UKHLS = UK Household Longitudinal Study; ONS = Office for National Statistics.

interactions can be estimated as smooth functions. Previous studies (e.g. Ellison et al., 2022; Ferrara & Vidoli, 2017; Reyes Santías et al., 2011) have found GAMs to improve efficiency and precision over standard, more restrictive methods when estimating complex underlying effects and interactions. In addition to their effective use of data and ability to learn about underlying patterns precisely by borrowing strength across covariate values, GAMs have many advantages in a predictive context. For example, GAMs are less prone to overfitting due to their efficient method of estimation and can therefore attain more robust out-of-sample fits (e.g. Berg, 2007; Lohmann & Ohliger, 2019). Their smooth framework also naturally extends to unobserved covariate values, conveniently linking the past and the future (e.g. Currie et al., 2004).

We include the covariates age, cohort, and time since last birth in our modelling process. Fertility trends across cohorts tend to be more stable than trends across calendar time (de Beer, 1985; Li & Wu, 2003), making the cohort approach taken in this paper appealing as a forecast dimension. We also consider two individual characteristics that are highly predictive of fertility behaviour. The first is highest educational qualification, a categorical variable with levels 'Less than General Certificate of Secondary Education (GCSE)'; 'GCSE'; 'Advanced (A) Level'; 'Degree'.[3] We impute values for the youngest women, many of whom are in education when interviewed and so have right-censored observations (see online supplementary Appendix B for an explanation of the imputation model). The second is the Human Development Index (HDI) of the country of birth, which we will refer to as 'birth HDI'. We use this measure since there is in general an association between the HDI and fertility rates (Myrskylä et al., 2009). We assign each woman an HDI using the 2018 values (United Nations Development Programme, 2019). We categorise the HDI values using the country groupings (United Nations Development Programme, 2018) which have levels 'low', 'medium', 'high', and 'very high' human development, and we also create a 'UK-born' category.

3.2 Model specification

We use discrete-time event history analysis with a logistic set-up (Steele, 2005). For a given parity $x \in \{0, 1, 2, 3+\}$, let Y_i be the sum of the n_i binary responses corresponding to the set of records sharing the ith covariate pattern, $i = 1, \ldots, N$, where N is the number of distinct covariate patterns observed in the sample of records. We assume a binomial distribution for Y_i with success probability r_i, i.e. $Y_i \sim \text{Binomial}(n_i, r_i)$. Note that r_i is the conditional probability of a birth event

[3] Note that each category includes equivalent qualifications, e.g. O Levels, CSEs, which existed when the older cohorts were in education. Generally, GCSEs are taken at the end of secondary schooling (around age 16) and A Levels at around age 18; A Levels are typically a requirement for university admission.

given the included covariates and x. The most complex possible GAM considered takes the form:

$$\begin{aligned}\text{logit}(r_i) = \mathbf{F}_i\boldsymbol{\theta} + f_A(a_i) + f_C(c_i) + f_T(t_i) + f_{AC}(a_i, c_i) + f_{AT}(a_i, t_i) + f_{CT}(c_i, t_i) \\ + f_{AQ}(a_i, q_i) + f_{CQ}(c_i, q_i) + f_{TQ}(t_i, q_i) + f_{AH}(a_i, h_i) + f_{CH}(c_i, h_i) + f_{TH}(t_i, h_i).\end{aligned} \tag{1}$$

The first term contains the fixed effects (the intercept; main effects of categorical variables and their interactions), where $\mathbf{F}_i$ is the ith row of the fixed effects model matrix with coefficients $\boldsymbol{\theta}$. The following terms are smooth functions of either the main effect of one variable or the interaction between two variables, with the variables involved indicated by the indexing [A = age; C = cohort; T = time since last birth top-coded at 11 years (available for $x \in \{1, 2, 3+\}$); Q= highest educational qualification; H= birth HDI]. We undertake a careful model selection process to determine which of these terms are included in the final model for each parity, described in Section 3.3. We note that the smooth interactions are two-dimensional (2D) functions unless one variable is categorical (Q or H), in which case the interaction is a one-dimensional (1D) smooth function of the continuous variable for each level of the categorical variable. The way in which we construct this latter kind of smooth interaction, as the combination of a global smooth function of the continuous variable, and level-specific functions with differing smoothness to capture the deviations from the global effect, allows the overall level-specific effects of the continuous variable to be shrunk towards the global effect. This set-up is akin to that of hierarchical generalised linear models, with smooth functions being pooled across groups as opposed to parameter estimates; by extension our GAM specification is also hierarchical (Pedersen et al., 2019).[4]

3.3 Model selection

To determine a chosen model for each parity, we fit and compare increasingly more complex forms of equation (1). Implementing a separate model selection process for each parity promotes parsimony and reduces the risk of overfitting by allowing each parity-specific model to be as complex as can be justified by the data. Referring to model components using the initials of the variables involved (i.e. U for the main effect of the variable U, UV for the interaction between the variables U and V), we begin by enforcing an A+C+AC baseline—that is, we include smooth main effects of age (A) and cohort (C) as well as their smooth interaction—so that the chosen models can fit sufficiently well to the population-level data. Our model selection process then involves the following steps:

1. For $x \in \{1, 2, 3+\}$, explore the inclusion of time since last birth (T) and its interactions with the other clocks (age and cohort).
2. Explore the inclusion of highest educational qualification (Q) and its interactions with the clocks.
3. Explore the inclusion of birth HDI (H) and its interaction with age.
4. Explore the inclusion of interactions between birth HDI and the remaining covariates.

As simpler alternatives to the qualification and HDI variables introduced in Section 3.1, in Steps 2–4 we also consider variants with two or more levels formed by combining adjacent categories. The selection process reflects our expectation that the clocks have the most explanatory power, followed by qualification and then HDI. We let qualification interact with the clocks because the literature supports at least an interaction with age (AQ), owing to greater postponement among more highly educated women (e.g. see Kravdal, 2001). There is also evidence for an AH interaction (e.g. see Waller et al., 2014), but for efficiency we only investigate further interactions involving HDI (i.e. proceed to Step 4) if the interaction with age provides a significant improvement in Step 3.

In each step, we take an approach akin to forward selection, comparing models using the Bayesian Information Criterion (BIC). In Step 1, we compare four models to the baseline model, obtained by adding T, T+AT, T+CT, and T+AT+CT. The model with the largest improvement in BIC becomes our new preferred model, which we then build on. Steps 2 and 3 proceed similarly, except that we

[4] We note the connection between the functional regression framework and hierarchical GAMs, namely that they can be viewed as function-on-scalar regression (Pedersen et al., 2019). We do not consider this classification further here, but see Pedersen et al. (2019) for additional information and key references.

consider all Q (Step 2) and H (Step 3) variants in tandem as at most one can be included. Then if we proceed to Step 4, the chosen variant of H (and possibly Q) is already determined. We fit the models in `R` (R Core Team, 2019) using the `mgcv` package (Wood, 2017). We take a P-spline approach (Eilers & Marx, 1996), expressing the 1D smooth functions as linear combinations of local B-spline basis functions and penalising the first differences of their coefficients to achieve suitable smoothness. The basis functions for 2D smooths are constructed by multiplying each of the univariate basis functions for the first variable with all of those for the second, penalising the first coefficient differences across both dimensions. Online supplementary Appendix C provides further details on the fitting process. As in Section 2, we weight the records using the survey weights.

We present the chosen models in Table 1, together with the form of their fitted probabilities. We refer to the chosen model for parity x as Mx, for $x \in \{0, 1, 2, 3+\}$. Note that Q subscripts indicate the number of levels in the particular variant: Q_4 is the original Q variable from Section 3.1; Q_3 has levels '< GCSE', 'GCSE/A Level' and 'Degree'; Q_2 has levels '< A Level' and 'At least A Level'. We first observe that M1–M3+ include time since last birth in addition to the A+C+AC baseline, demonstrating its significance. The models for higher parities are less complex, supported by the presence of qualification in all models barring M3+, AQ only in M0 and M1, and Q_4 only in M0. This simplification is likely partially accounted for by the fewer birth events at higher parities (Ellison et al., 2022). The fact that the two selected interactions involve age highlights its explanatory power. Lastly, the absence of HDI could be partly due to its consideration after qualification in our selection process. Indeed, for each parity at least two of the five models with the lowest BIC include HDI, evidencing its explanatory ability.

4 Incorporating the population-level data

4.1 Introduction

In Section 1, we discussed the potential for bias in the ONS age–parity-specific fertility rates and summarised two studies that account for biases in the population-level data. Their methodologies require sufficient knowledge about the precise mechanism through which the bias arises. In our case such details are not available, and the mechanism is likely to be highly complex and parity-specific. Furthermore, the aim of this paper is to provide a general methodology that can be applied to a range of datasets. Therefore, we model the ONS data as unbiased. For a given parity x, age a, and cohort c, let $Y_{a,c}$ be the number of births and $n_{a,c}$ the exposure. As with the UKHLS data in Section 3.2, we make a binomial assumption for the ONS birth counts, i.e. $Y_{a,c} \sim \text{Binomial}(n_{a,c}, r_{a,c})$. Consistent with Table 1, we define the success probability $r_{a,c} \equiv \hat{P}_x(\text{birth}|a, c)$, $x \in \{0, 1, 2, 3+\}$ as the generic form of these fitted probabilities.

In order to integrate the ONS data into our chosen GAMs, we express each $\hat{P}_x(\text{birth}|a, c)$ in terms of its corresponding fitted probability in Table 1, i.e. marginalise over the additional qualification (Q) and/or time since last birth (T) covariates. Such an approach is reminiscent of the general method used by Handcock et al. (2005) and Rendall et al. (2009) to construct their constraint functions. To this end, we can write:

$$\hat{P}_0(\text{birth}|a, c) = \sum_{q_4} \hat{P}_0(\text{birth}|a, c, q_4)\hat{P}_0(q_4|a, c); \tag{2}$$

$$\begin{aligned}\hat{P}_1(\text{birth}|a, c) &= \sum_{t,q_2} \hat{P}_1(\text{birth}|a, c, t, q_2)\hat{P}_1(t, q_2|a, c) \\ &= \sum_{t,q_2} \hat{P}_1(\text{birth}|a, c, t, q_2)\hat{P}_1(q_2|a, c)\hat{P}_1(t|a, c, q_2);\end{aligned} \tag{3}$$

$$\begin{aligned}\hat{P}_2(\text{birth}|a, c) &= \sum_{t,q_3} \hat{P}_2(\text{birth}|a, c, t, q_3)\hat{P}_2(t, q_3|a, c) \\ &= \sum_{t,q_3} \hat{P}_2(\text{birth}|a, c, t, q_3)\hat{P}_2(q_3|a, c)\hat{P}_2(t|a, c, q_3);\end{aligned} \tag{4}$$

$$\hat{P}_{3+}(\text{birth}|a, c) = \sum_{t} \hat{P}_{3+}(\text{birth}|a, c, t)\hat{P}_{3+}(t|a, c). \tag{5}$$

Table 1. Chosen models for parities 0–3+ (M0–M3+) with the generic form of their fitted probabilities

Model	Model description	Fitted probability
M0	$A+C+AC+Q_4+AQ_4$	$\hat{P}_0(\text{birth}\|a, c, q_4)$
M1	$A+C+AC+T+Q_2+AQ_2$	$\hat{P}_1(\text{birth}\|a, c, t, q_2)$
M2	$A+C+AC+T+Q_3$	$\hat{P}_2(\text{birth}\|a, c, t, q_3)$
M3+	$A+C+AC+T$	$\hat{P}_{3+}(\text{birth}\|a, c, t)$

Note. A= age; C = cohort; Q= highest educational qualification (see text for variant definitions); T = time since last birth. Single letters are main effects, pairs of letters are interaction effects.

For each parity, we have now expressed the marginalised probability $\hat{P}_x(\text{birth}|a, c)$ as a weighted sum of the original GAM probabilities across all combinations of the corresponding Q and/or T values. The weights are (products of) conditional probabilities of these covariate values given other variables in the model. We classify the covariate probabilities by whether they model qualification given age and cohort, or time since last birth given age, cohort, and qualification (note that dependence on qualification is only present for parities 1 and 2). In order to determine appropriate models for the covariate probabilities within these two groups, we require parity-specific counts by the variables involved for England and Wales. We obtain these from our sample of UKHLS records. We note that in order to marginalise the projections, it is necessary for these models to provide projected values for the unobserved age–cohort combinations. In Sections 4.2 and 4.3, we detail the modelling processes for the two groups.

4.2 Modelling qualification given age and cohort

Here, we model the dependence of Q_4, Q_2, and Q_3 (variant definitions given in Section 3.3) on age and cohort for parities 0, 1, and 2, respectively. We omit records for the post-1982 cohorts as their qualification values are imputed, and it is undesirable for our modelling decisions to be imputation-dependent. We determine suitable models for each parity which we then extrapolate, giving smooth fitted age–cohort surfaces for all the desired age–cohort combinations, i.e. for each of the 1945–2003 cohorts from ages 15 to 44 (see Section 2). We note that some of these combinations, such as the 1995 cohort at age 40, are in the future.

We briefly describe our modelling approach and selection process, with further details given in online supplementary Appendix E. We model the probability of belonging to each qualification category using Bayesian multinomial logistic regression, with the first level as reference category. We experiment with various specifications of the linear predictors and fit the models in `R` using the `rstan` software package (Stan Development Team, 2019), which implements the efficient Hamiltonian Monte Carlo methodology. We compare models using the BIC and perform a rough assessment of fit using a Pearson chi-squared statistic, selecting the model with the lowest BIC and a sufficiently close fit for each parity. In order to visualise the results of the model selection process, in Figure 2 we plot the fitted (posterior mean) probabilities of belonging to each qualification category for a given age, cohort and parity. We achieve smoothness across age by constraining each series of age-specific parameters using a first-order random walk prior—this borrowing of strength is especially important for parities 1 and 2, where we have small exposures at the youngest ages and therefore considerable uncertainty about the corresponding parameters. Comparing the portions to the left of the dividing lines (the age–cohort combinations used in model fitting) with the observed proportions (not shown here), they exhibit the same features and trends but with greater smoothness across both dimensions. This facilitates a plausible and appropriate extrapolation to future age–cohort combinations (the portions to the right of the dividing lines).

4.3 Modelling time since last birth given age, cohort, and qualification

To model the dependence of time since last birth (T) on age and cohort for parity 3+, and additionally on qualification for parities 1 and 2, we take a similar approach to Section 4.2.[5] We present the

[5] Again we omit records corresponding to the post-1982 cohorts for parities 1 and 2, where we have dependence on qualification.

fitted (posterior mean) probabilities of belonging to each time since last birth category for a given age, qualification level, and parity in Figure 3, and highlight some key features of the surface before summarising the modelling approach (online supplementary Appendix F gives further details). First, the number of observable time since last birth categories increases from 3 at age 15 to our top-coded maximum of 11 for ages 23+.[6] Second, there are opposing trends at the youngest and oldest ages, the proportions peaking at $T = 1$ and $T = 11$, respectively. This is intuitive because births to young women are likely to have occurred very recently, while at older ages the opposite is true. Third, we note that each diagonal line of cells parallel to the 'step' line corresponds to a different age at last birth event, increasing from 12 to 43. The approximate parallelograms of higher density in each surface thus represent the commonest ages at entry into the particular parity (or possible re-entry in the case of parity 3+). Further disaggregation by cohort resulted in surfaces exhibiting similar patterns. However, the smaller exposures led to increased parameter uncertainty especially for the younger cohorts. We therefore ignore dependence on cohort, enabling maximal borrowing of strength and minimal extrapolation.

As the number of observable time since last birth categories varies with age, we specify a separate multinomial model for each value from 3 to 11. However, we share T-specific parameters across age (and possibly qualification) by fitting the models simultaneously to the time since last birth counts for a given parity. We fit Bayesian models with various forms of the linear predictors, and perform model comparison and informal fit assessment as in Section 4.2. Again we employ a first-order random walk prior, this time to smooth the age-at-entry-specific parameters, reducing their uncertainty particularly at younger ages. The correspondence of the fitted probabilities in Figure 3 with the observed proportions (not shown here) is closest for older ages at entry, but the fit at younger ages looks plausible given that we aimed to obtain a smooth surface across all observable combinations of age and time since last birth.

4.4 Integrating the data sources

To incorporate the population-level data into our model for the individual-level data, we combine our chosen GAMs (Section 3.3) with the elements developed in Section 4. Using the notation of Section 3.2, for a given parity x and the ith covariate pattern we let n_i^w and $\hat{r}_i = y_i/n_i$ be the weighted number of records and the unweighted proportion of successes respectively. Then we express the binomial log-likelihood contribution for the UKHLS (individual-level) data, ℓ_{INDIV}, as:

$$\ell_{\text{INDIV}} = \sum_{i=1}^{N} n_i^w \{\hat{r}_i \log r_i + (1 - \hat{r}_i) \log (1 - r_i)\}. \tag{6}$$

This adjusts the weighting method slightly from Section 3.3 to provide the integer number of trials and successes required by the `Stan` software (Stan Development Team, 2019), which we use for model fitting.[7] Using the notation of Section 4.1, the equivalent quantity for the ONS (population-level) data, ℓ_{POP}, is:

$$\ell_{\text{POP}} = \sum_{a,c} n_{a,c} \{\hat{r}_{a,c} \log r_{a,c} + (1 - \hat{r}_{a,c}) \log (1 - r_{a,c})\}, \tag{7}$$

where $\hat{r}_{a,c} = y_{a,c}/n_{a,c}$ is the observed proportion of successes. Note that the summation is over all observed age–cohort combinations in the ONS data.

If we were to simply maximise the sum of the likelihoods in equations (6) and (7), we would essentially be combining the UKHLS and ONS populations; consequently, the information contained in the ONS data would dominate our overall inference due to its significantly larger exposures. To balance the sources of information we take a logarithmic pooling approach. Logarithmic pooling of prior distributions is commonly used to combine the opinions of experts in Bayesian

[6] This is because we censored any births to women aged less than 12 at the time of the birth, treating their dates as missing.

[7] In this way, we still use the weighted number of trials but replace the weighted proportion of successes with the unweighted proportion. This amounts to scaling the trial and success counts by the same factor, namely the quotient of the weighted and unweighted number of trials.

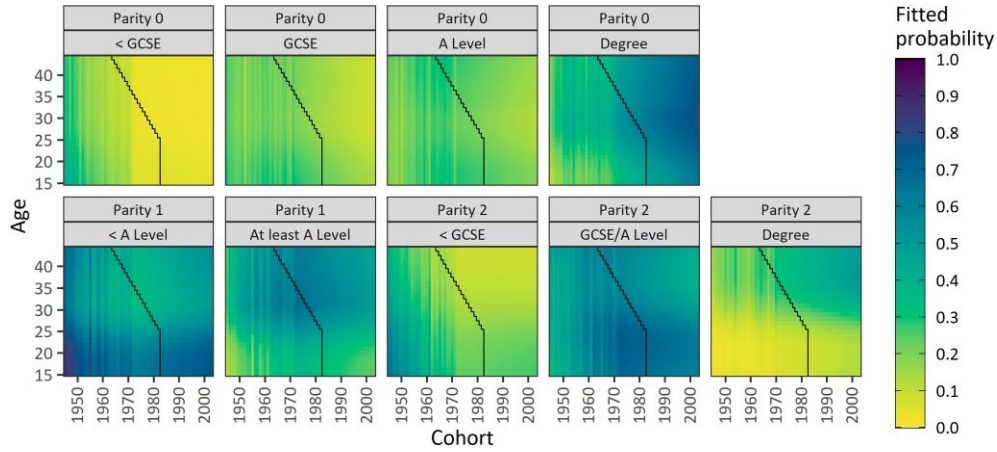

Figure 2. Parity-specific Lexis surfaces of the fitted probabilities corresponding to the final models for each qualification category and a given age and cohort. Qualification categories correspond to the variant included in the particular chosen model in Table 1. The dividing lines indicate the age–cohort combinations used in model fitting (the 1945–1982 cohorts, observed up to 2007); to the right of these lines the probabilities are extrapolations.

analyses (e.g. see Carvalho et al., 2023; Poole & Raftery, 2000). This method has also been applied to likelihoods to integrate information from multiple independent data sources and subsequently improve predictive performance (Fletcher et al., 2019; Hu & Zidek, 2002). Taking this latter implementation, we introduce a weight $w \in [0, 1]$ that gives us control over the joining of the two populations. Weighting the UKHLS and ONS contributions in equations (6) and (7) by $(1 - w)$ and w, respectively, we then obtain the following combined log-likelihood contribution, ℓ_{comb}:

$$\ell_{\text{comb}} = (1 - w) \times \ell_{\text{INDIV}} + w \times \ell_{\text{POP}}. \tag{8}$$

In this way, with the proportion of successes remaining unchanged, we are able to downweight each of the UKHLS and ONS exposures by $(1 - w)$ and w respectively. We observe that setting $w = 0$ corresponds to putting full weight on ℓ_{INDIV} and no weight on ℓ_{POP}, thereby fitting a UKHLS-only model; conversely, setting $w = 1$ corresponds to an ONS-only model. For $w \in (0, 1)$, the combined population is a weighted average of the UKHLS and ONS populations.

A key challenge of this approach is choosing the value of w—to inform this, we derive an approximate interpretation. In Section 2, we noted that the ONS exposures are around 1,000 times greater than the corresponding UKHLS exposures. If, for convenience, we approximate the ratio of the UKHLS to the ONS exposures by 1:999, and believe a priori that the relative importance of the two datasets to overall inference is INDIV:POP, it can be easily shown that it is appropriate to set $w = \text{POP}/(999 \times \text{INDIV} + \text{POP})$. This does however assume equal overdispersion in the two datasets, which may be unrealistic. In Table 2, we present various ratios INDIV:POP, together with the corresponding value of w. Starting from a 9:1 ratio, we increase the ONS contribution up to a 1:9 ratio, generating weights between 0.0001 and 0.009. This demonstrates that due to the large exposures in the ONS data, an a priori belief that the datasets are equally important only requires a very small value of w (0.001). We note that the derivation, and by extension the correspondence between INDIV:POP and w, is not precise—it just gives a rough indication of how the value of w can be interpreted.

4.5 Fitting the Bayesian parity-specific fertility projection model

In this subsection, we summarise the fitting process for our Bayesian parity-specific fertility projection model. First, as in Sections 4.2 and 4.3, we use the `rstan` software package (Stan Development Team, 2019). We therefore need to formulate Bayesian GAMs that incorporate the `mgcv` set-up in terms of basis functions and penalisation (see Section 3.3 and online supplementary Appendix C), with some adaptations to suit our proposed projection model.[8]

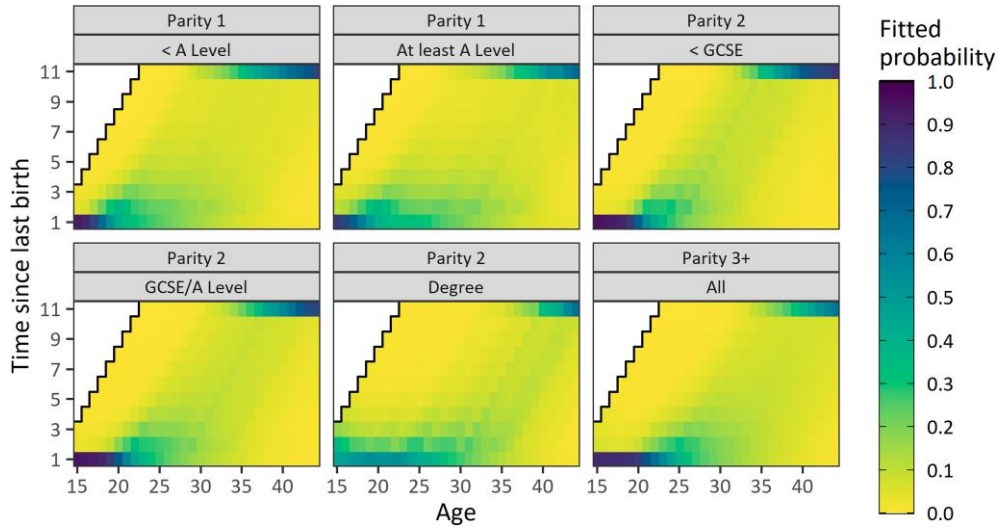

Figure 3. Parity-specific Lexis surfaces of the fitted probabilities corresponding to the final models for each time since last birth category and a given age and qualification category. Qualification categories correspond to the variant included in the particular chosen model in Table 1. The empty cells in the top-left corner of each plot indicate age–time since last birth combinations that are not observed, as they correspond to ages at birth younger than 12.

We do this by combining the likelihood in equation (8) with priors on the parameters controlling the smoothness of each of the functions in our particular chosen GAM (see Table 1) to express our belief that smoother functions are more likely.

For ease of explanation, we describe the construction of the prior for a general smooth function that can either be 1D or 2D (see Section 3.2). Following from the description of the P-spline approach given in Section 3.3, we model this function as the product $\mathbf{M}\boldsymbol{\mu}$, where $\mathbf{M}$ is the matrix of basis functions and $\boldsymbol{\mu}$ the vector of basis function coefficients.

Taking a 1D smooth function first, we let $\mathbf{S}$ be the penalty matrix such that $\boldsymbol{\mu}'\mathbf{S}\boldsymbol{\mu}$ gives the sum of the squared first coefficient differences to be penalised (see Section 3.3). Lastly, we define the smoothing parameter τ, which controls smoothness such that smaller values of τ impose greater smoothing. Following Umlauf et al. (2018), we then specify the prior for $\boldsymbol{\mu}$, conditional on τ, as

$$f(\boldsymbol{\mu}|\tau) \propto \left|\frac{\mathbf{S}}{\tau}\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}'\left\{\frac{\mathbf{S}}{\tau}\right\}\boldsymbol{\mu}\right). \tag{9}$$

For a 2D smooth function, the construction of the prior extends naturally to accommodate the fact that the coefficient differences are penalised across two dimensions rather than one (Section 3.3). Consequently, we have two penalty matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$, and two smoothing parameters, τ_1 and τ_2. We specify the resulting prior as:

$$f(\boldsymbol{\mu}|\tau_1, \tau_2) \propto \left|\frac{\mathbf{S}_1}{\tau_1} + \frac{\mathbf{S}_2}{\tau_2}\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}'\left\{\frac{\mathbf{S}_1}{\tau_1} + \frac{\mathbf{S}_2}{\tau_2}\right\}\boldsymbol{\mu}\right). \tag{10}$$

We specify a weakly informative $N^+(0, 10^2)$ half-normal prior on smoothing parameters τ to give the data sufficient freedom to determine appropriate smoothness. We specify $N(0, 10^2)$ priors on all other parameters.

[8] We adapt the `mgcv` basis functions in two ways: (1) We amend our basis functions for the age and cohort smooths so that they cover the required ranges (15–44 and 1945–2003, respectively). (2) We retain the linearly dependent basis functions for the 2D smooths, which are removed in the `mgcv` fitting process (see Section C.3 of the online supplementary Appendix C), to obtain more appropriate forecast intervals.

Table 2. Various a priori beliefs about the relative importance of the UKHLS and ONS datasets INDIV:POP, with the corresponding required weight w on the ONS exposures (given to 1 significant figure)

INDIV:POP	w
9:1	$1/8{,}992 \approx 0.0001$
2:1	$1/1{,}999 \approx 0.0005$
1:1	$1/1{,}000 = 0.001$
1:2	$2/1{,}001 \approx 0.002$
1:9	$9/1{,}008 \approx 0.009$

Then, for each parity we integrate the `Stan` code for the chosen GAM (Section 3.3) and the relevant covariate models (Sections 4.2 and 4.3), allowing simultaneous fitting. We can therefore access the resulting probabilities and marginalise over qualification and/or time since last birth to obtain the $r_{a,c}$'s (Section 4.1). Finally, we adjust the log-likelihood contribution as in equation (8) to incorporate the ONS data. We refer to these models, which combine the data sources using a particular weight $w \in [0, 1]$, as 'integrated models'. We perform 1,000 warmup and 1,000 retained iterations for each model, which takes approximately 3, 9.5, 11, and 3 h on average for parities 0, 1, 2, and 3+, respectively, on a 2.7 GHz Intel Core i7 Windows machine. This correlates strongly with the number of covariate patterns N in the UKHLS dataset (Section 3.2): while parities 0 and 3+ include only one of qualification and time since last birth in their chosen models (Table 1), parities 1 and 2 include both and therefore have considerably larger values of N.

5 Results

5.1 Initial examination of the effect of varying the pooling weight

For each parity, we fit a Bayesian version of the UKHLS-only GAM chosen in Section 3.3, and integrated models for the INDIV:POP ratios proposed in Table 2. We begin by investigating the effect of varying INDIV:POP on the model fits, forecasts, and their plausibility. In Figure 4, we plot the posterior means of the marginalised probabilities, i.e. the $r_{a,c}$'s (see Section 4.1) for all age–cohort combinations and ascending w, with the observed ONS rates in the bottom row. Consequently the figure presents a spectrum of models, the ONS contribution increasing as we move downwards. For convenience, we refer to the integrated models using the odds notation in Table 2. We reiterate that these should not be overinterpreted, and are simply being used as they carry more meaning than w itself.

For parity 0, comparing the projections from the 1:0 (UKHLS-only) model to those from the 9:1, 2:1, and 1:1 models, we see that incorporating the ONS data reverses the increases forecast by the former to increasingly slower declines, which are quickest at the youngest ages. This is driven by the additional observations from the 1993–2003 cohorts in the ONS data (see Section 2). In particular, the diminishing rates of early motherhood only start to occur for these more recent cohorts, so the 1:0 model forecasts could not be influenced by this. As the ONS contribution increases further to the 1:9 model, we see stronger recuperation of first births at older ages; this is likely to be due to the greater emphasis on recent declines. Comparing the analogous plots for parity 1, the differences, albeit less drastic than parity 0, are again most noticeable for the youngest cohorts. We also see that the fitted probabilities at around age 20 for the oldest cohorts reduce slightly to align with the ONS rates. Increasing w shifts the projected curves for the youngest cohorts to the left, leading to increases at younger ages and faster decreases at older ages. This is likely an effect of overfitting to the little ONS data available for these cohorts, and can be seen most clearly in the 1:9 model forecast.

The impact of introducing the ONS data is significant for parity 2, the shape of the curves from ages 15 to 22 transforming from a gradually increasing to sharply declining trend from a much higher level. A dramatic change at these young ages is unsurprising given the erratic nature of their observed rates due to low exposures (see Section 2). Regarding the forecasts, the curves for the youngest cohorts are close for small w, however with larger w they decline with quickening speed;

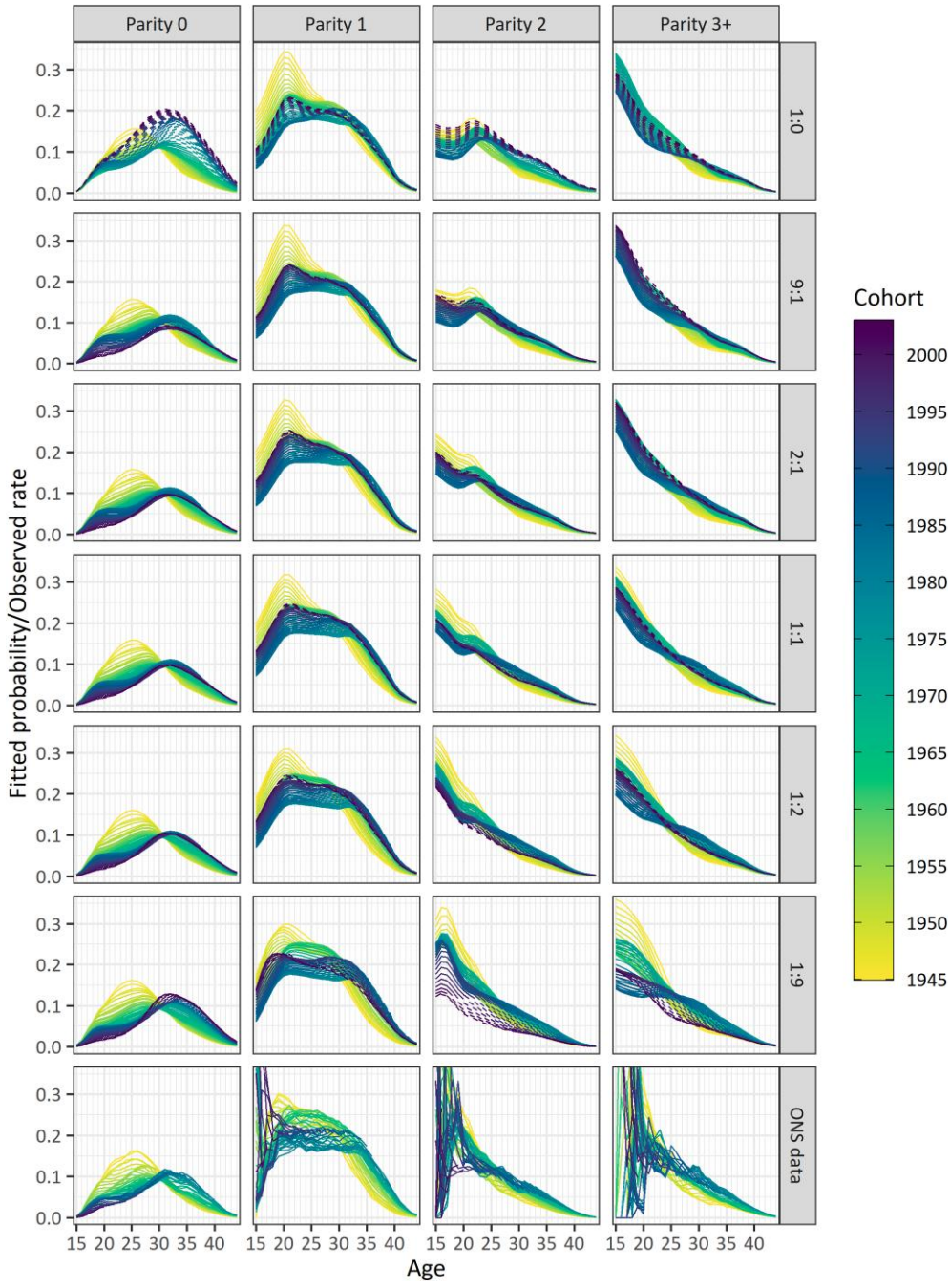

Figure 4. Fitted (posterior mean) probabilities of a birth event given age, cohort, and parity for various integrated models, and the observed parity-specific fertility rates from ONS (2020). Dashed lines indicate forecasts, i.e. age–cohort combinations after the most recent year of observation across the included datasets.

this presents even stronger evidence that the need to fit closely to the ONS data is dominating the need for smoothness and hence overfitting is occurring. Lastly, the parity 3+ forecasts only change slightly for the youngest cohorts between the 1:0 and 1:1 models. Moving towards the 1:9 model there is greater disparity across the curves at the youngest ages, and a gradual lowering of the curves for the most recent cohorts.

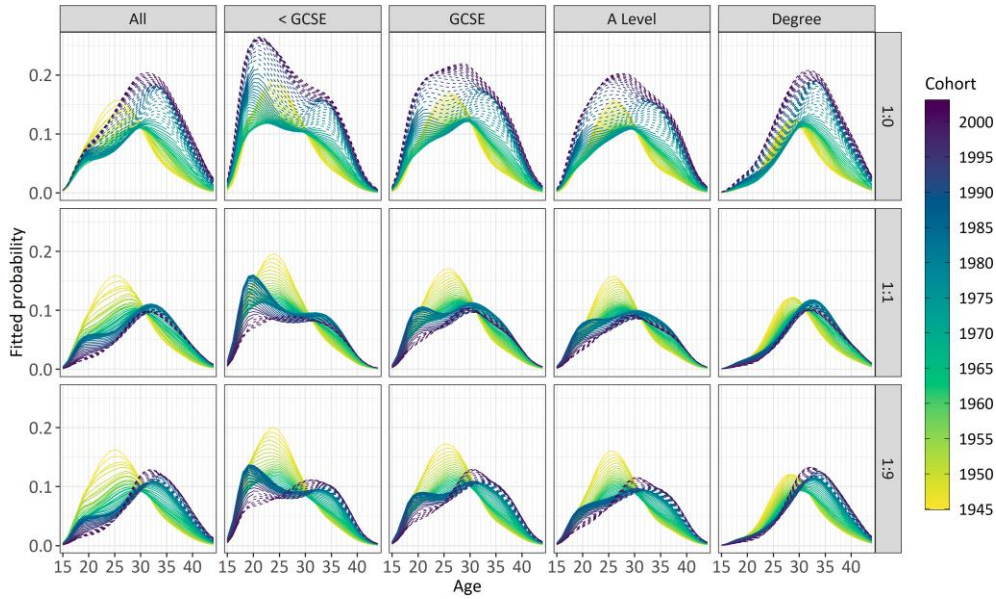

Figure 5. Fitted (posterior mean) probabilities of a birth event given age, cohort, highest educational qualification and being in parity 0 for various integrated models. Dashed lines indicate forecasts, i.e. age–cohort combinations after the most recent year of observation across the included datasets.

In conclusion, we see that even the 9:1 model considerably adjusts the forecasts for the youngest cohorts, aligning the fits with the ONS rates reasonably precisely. Increasing the ONS contribution up to the 1:9 model improves this alignment but also overfits to the erratic rates at young ages, leading to unrealistic declines forecast for parities 1 and above. For convenience, in the following subsections we will concentrate on a subset of these models, namely the 1:0, 1:1, and 1:9 models, to contrast examples with high UKHLS, roughly equal, and high ONS contributions, respectively.

5.2 Unmarginalised results

In this section, we present the probabilities directly modelled by our chosen GAMs from Section 3.3, i.e. those depending on qualification (Q) and/or time since last birth (T) in addition to age and cohort. We focus on parities 0 and 1 as the presence of age–qualification interactions means that when plotting the probabilities by age for a given cohort, they exhibit more interesting features than a simple change in level. Starting with parity 0, in Figure 5 we plot the posterior mean probabilities for the 1:0, 1:1, and 1:9 models by age and cohort for each qualification level, with the marginalised probabilities from Figure 4 in the first column. There is strong evidence of postponement across qualification: for example, the curves for the mid-1980s cohorts are strongly bimodal with peaks at younger and older ages for the lower qualification levels, but unimodal with one higher peak at older ages for the 'Degree' category. A similarly changing age pattern is identified and discussed extensively in Ellison et al. (2022).

As in Section 5.1, we see the reigning in and reversal of the optimistic and explosive 1:0 model forecasts upon incorporating the ONS data. However, being able to plot the qualification-specific forecasts gives us an additional insight into the mechanism through which these changes can be realised. For example, the forecasted declines in teenage fertility mostly occur to those with lower levels of education, whose curves change from strongly bimodal to unimodal for the youngest cohorts (Figure 5). In this way, the model predicts a convergence to a late childbearing pattern across the qualification categories. The ability of our integration method to retain qualification and time since last birth, and hence allow us to improve forecast plausibility for population subgroups determined by these variables despite the ONS data not explicitly informing about them, is a key strength of our approach.

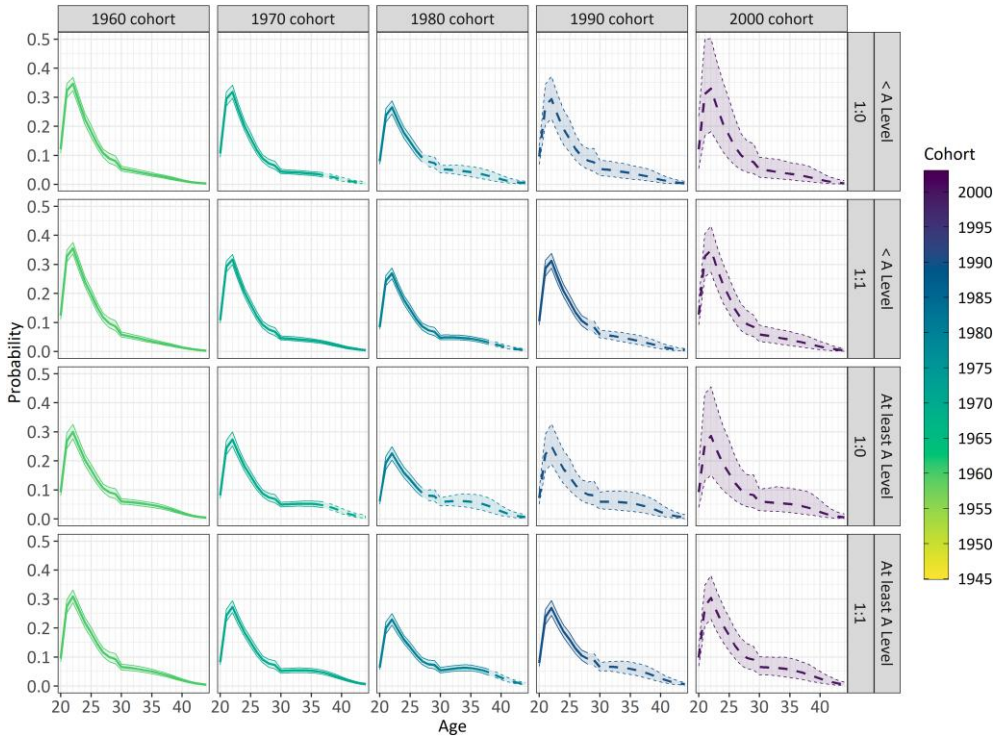

Figure 6. Fitted (posterior mean) probabilities of a birth event given age, cohort, highest educational qualification, time since last birth and being in parity 1, with age at first birth fixed at 19 years (thick lines), and their corresponding 95% credible intervals (thin lines), for various integrated models. Dashed lines indicate forecasts, i.e. age–cohort combinations after the most recent year of observation across the included datasets.

Next, we present the parity 1 results, two key differences being that the qualification variant, Q_2, has two categories ('< A Level' and 'At least A Level'), and time since last birth is included. In Figure 6, for the 1:0 and 1:1 models we plot the posterior mean birth probabilities and their corresponding 95% credible intervals (CIs), for selected cohorts and an age at first birth of 19 years, as this is the most common value in our sample. In this way, $A = 20$ is equivalent to $T = 1$, $A = 21$ to $T = 2$, and $A \geq 30$ to $T = 11$, and so the effects of age and time since last birth are confounded. The likelihood of a second birth peaks sharply at 2–3 years after the first, before declining steeply. We still see postponement, here through the decrease and increase in level at younger and older ages respectively for the 'At least A Level' category. The effects of incorporating the ONS data identified in Section 5.1 persist, albeit less noticeably due to the confounding and truncated age range. Regarding the CIs, we observe a significant reduction in uncertainty through integrating the ONS data; this further illustrates the benefits of combining the datasets for inferring about variables that are unavailable at the population level. The CIs also indicate that the peak around $A = 22$ is not so severe for cohorts currently younger than this; for example, the intervals for the 2000 cohort under the 1:1 model allow for reasonable flexibility in its location, height, and sharpness.

5.3 Marginalised results

Next, we concentrate on the marginalised probabilities, i.e. those depending on age and cohort only. We plot the posterior means in Section 5.1 so here, for the 1:1 model and each parity, in Figure 7 we present the 95% CIs for the underlying probabilities for selected cohorts. We also present results from fitting the 1:1 model using the ONS data available in 2013 rather than 2018, the five years of holdout data allowing us to assess forecast sensitivity. We overlay the ONS rate estimates for comparison, but we do not expect 95% of them to lie within the CIs; this is mainly because the fit is also influenced by the UKHLS data. We note that the intervals do not include

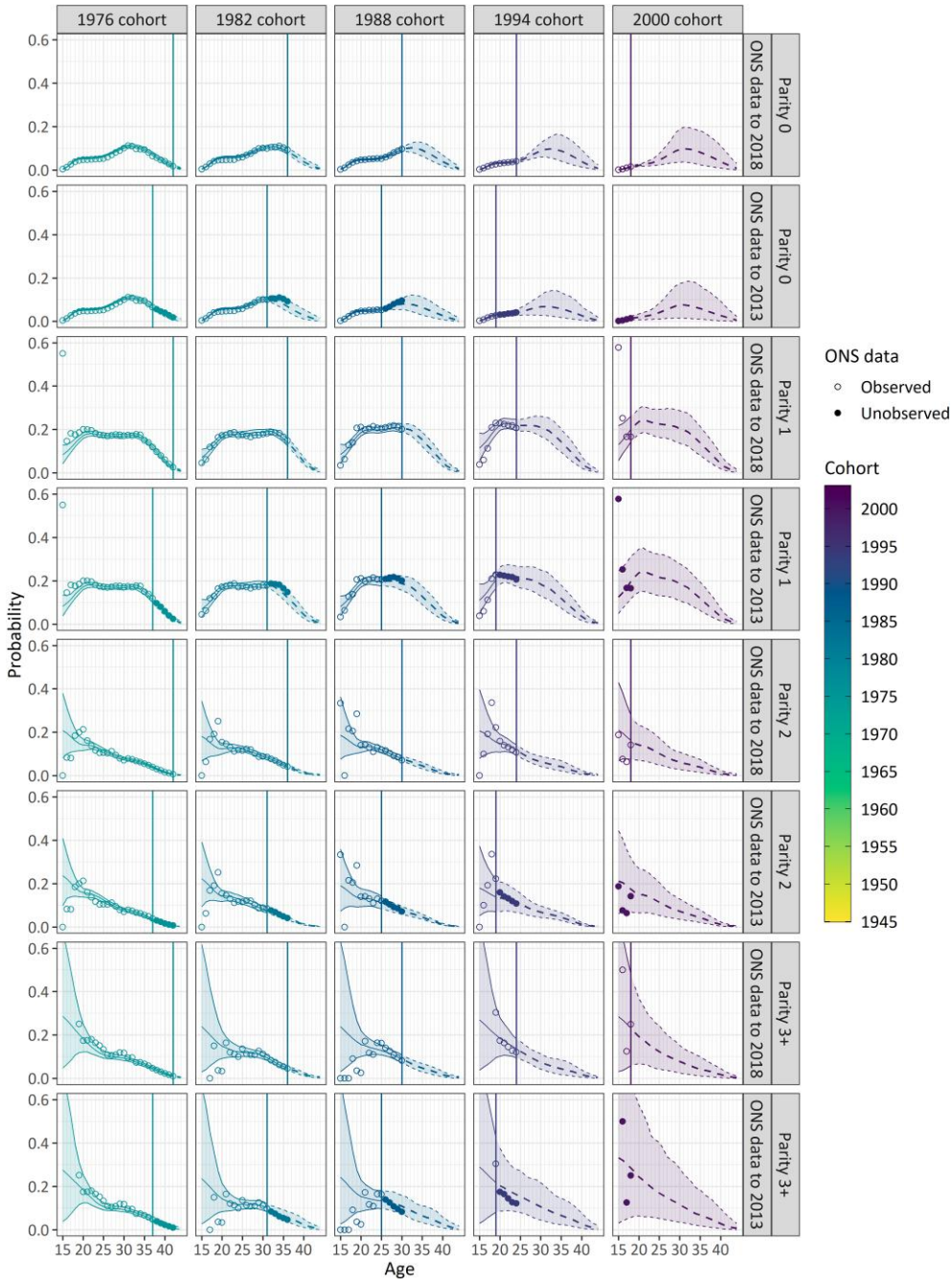

Figure 7. Fitted (posterior mean) probabilities of a birth event given age, cohort and parity (thick lines), and 95% credible intervals (thin lines), for the 1:1 model fitted using ONS data to 2018 and 2013. 'Observed' and 'Unobserved' refer, respectively, to the data points included in, and excluded from, model fitting. Dashed lines indicate forecasts, i.e. age–cohort combinations after the most recent year of observation across the included datasets; vertical lines correspond to this most recent year of observation. ONS = Office for National Statistics.

predictive uncertainty, which would widen them slightly where exposures are small, i.e. for the youngest ages at parities 1 and above.

For parity 0, we plot the results for the full and truncated datasets respectively in rows 1 and 2 of Figure 7. Each vertical line indicates the last observed age, with points to the right excluded from

model fitting (i.e. unobserved). We see that truncating the dataset leads to severe declines in the fertility level across all ages for the younger cohorts, whereas with the full dataset the decreases at young ages are partially offset by a strong recovery predicted at older ages (see also Figure 4). Despite this, the unobserved rates tend to lie in the intervals, albeit towards the upper bound, which is expected given the more negative forecasts. This suggests that a tighter prior on the smoothing parameter for the cohort main effect may be necessary to incorporate our prior belief that cohort changes are relatively slow, thereby reducing the risk of quickly emerging trends at young ages being extrapolated across the age range. Although the change in level makes it difficult to compare the widths of the CIs for the youngest cohorts, for slightly older cohorts with more observed data (e.g. the 1980s cohorts) the intervals are noticeably wider with the truncated dataset, as anticipated.

For parity 1 (rows 3 and 4 of Figure 7), the posterior means change less dramatically with the truncated dataset, and there is a clear increase in forecast uncertainty for all cohorts. However, we see less stability across cohorts at older ages, with reasonably strong decreases in level for the youngest cohorts (e.g. the 2000 cohort) at ages above 30. The CIs perform slightly better compared to those for parity 0, the unobserved points occupying different regions of the intervals; they are also wide enough to account for the erratic observations at young ages. For parities 2 and 3+ (rows 5–8 of Figure 7) the widening of the intervals is even more striking, with the truncated dataset giving rise to increases for the youngest cohorts at younger ages (most notably the 2000 cohort), which are particularly steep for parity 3+ at ages 15–30. While these optimistic forecasts appear to be inappropriate for parity 2, with the unobserved points near the lower bounds of the CIs, they look to be more justifiable for parity 3+.

5.4 Aggregate results

In this final subsection, we additionally marginalise over parity to obtain measures on the aggregate level, namely ASFRs and completed family size (CFS). For a particular cohort c, the ASFR at age a is calculated by dividing the total number of births by the total number of women; summing these ASFRs across the reproductive age range then gives the CFS, which represents the average number of children born to women in cohort c. Inspired partly by the methodology of Smallwood (2002), we describe our approach for a given model and iteration of its marginalised probabilities. We take the age–cohort exposures as known, using midyear population estimates from ONS (2020) up to 2018, and the ONS 2018-based National Population Projections (NPPs) thereafter (ONS, 2019). For a given cohort, we first assume that all women are childless at the youngest reproductive age considered (15 in our case). Next, at each age we sample order-specific birth counts from binomial distributions with number of trials equal to the estimated parity-specific exposures, and success probabilities equal to the corresponding marginalised probabilities. We use these to update the number of women exposed at each parity at the next age, which we then scale so that they sum to the appropriate midyear population estimate.

For each age, aggregating the births and dividing by the midyear population estimate gives a sample of the ASFR, and summing the ASFRs over age gives a sample of the CFS. Repeating this process for each iteration and cohort gives us 1,000 samples of each ASFR and CFS, from which we calculate posterior means and 95% CIs. In Figure 8, we present these quantities for the ASFRs of selected cohorts, for the 1:1 model with truncated and full ONS datasets (see Section 5.3), and the 1:9 model. Similarly to Figure 7, we overlay the ONS rate estimates. However, as these are direct estimates and our intervals incorporate predictive uncertainty (through the binomial variability), we can assess the fit and forecast performance of our integrated models more reliably.

The fits of the 1:1 models are reasonably close, with some overprediction for the 1976 and 1982 cohorts at ages 20–30. This discrepancy diminishes with more weight on the ONS data in the 1:9 model. As in Section 5.3, the forecast intervals noticeably widen upon truncating the ONS dataset; they include most of the unobserved data points, but perform poorly for the 1982 cohort. The 1:9 model CIs are slightly narrower than those for the corresponding 1:1 model, owing to the larger combined population in the likelihood (see Section 4.4). The last row displays the projected ASFRs under the principal variant of the 2018-based NPPs. We observe that these are considerably higher than the 1:9 model posterior means at ages 20–25 (e.g. the 2000 cohort) and 30–35 (e.g. the 1994 cohort).

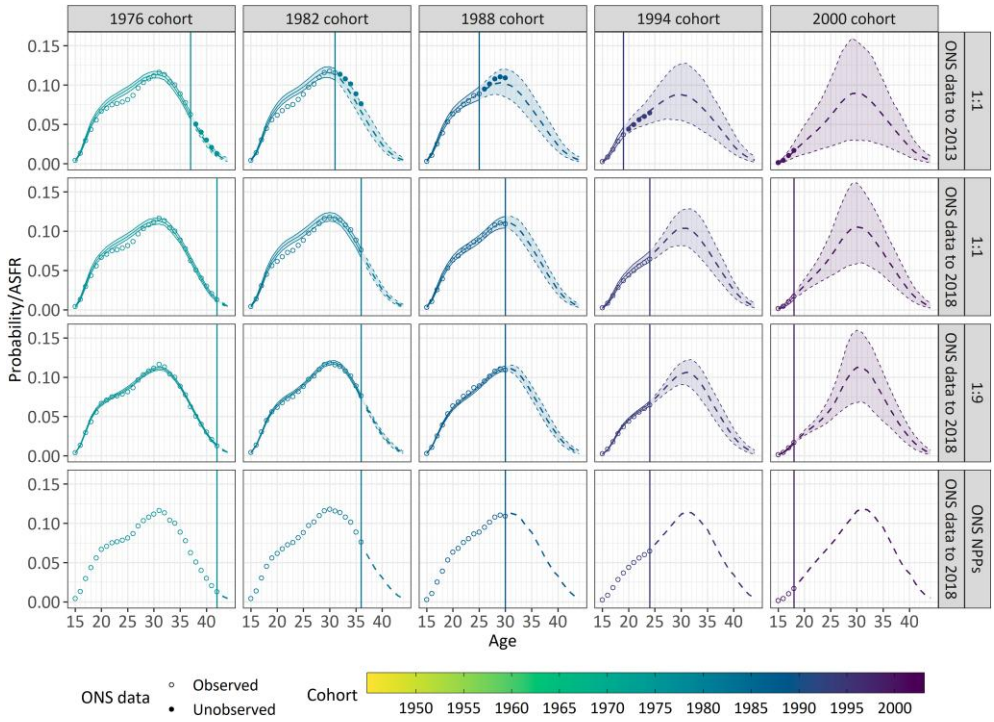

Figure 8. Fitted (posterior mean) probabilities of a birth event given age and cohort (thick lines), and 95% credible intervals (thin lines), for various integrated models fitted using ONS data to 2013 and 2018, as well as the ASFRs from the principal variant of the ONS 2018-based NPPs (bottom row). 'Observed' and 'Unobserved' refer respectively to the data points included in, and excluded from, model fitting. Dashed lines indicate forecasts, i.e. age–cohort combinations after the most recent year of observation across the included datasets; vertical lines correspond to this most recent year of observation. ONS = Office for National Statistics; ASFRs = age-specific fertility rates; NPPs = National Population Projections.

Figure 9 plots the 50% and 90% CIs for the CFS under the 1:1 and 1:9 models fitted to the full ONS dataset, overlaying the observed values. As in Figure 8, the 1:9 model fit is closer and the forecast CIs are narrower. We additionally overlay the corresponding CIs for the Bayesian cohort fertility forecasting model of Schmertmann et al. (2014a), which was found to be one of the best-performing methods in a recent comparative paper (Bohk-Ewald et al., 2018; Ellison et al., 2020). We fit the model using the code provided on the accompanying website (Schmertmann et al., 2014b), with minor modifications to update the input data. Comparing with the 1:9 model (right panel), we see that both models predict broadly similar trends in the CFS with comparable levels of uncertainty; however, the decline is projected to begin slightly later under the model of Schmertmann et al. (2014a) and so the intervals remain at higher levels. We also overlay CFS projections derived from the principal, low fertility and high fertility variants of the 2018-based NPPs. They represent three distinct scenarios rather than a probability interval so we cannot perform a direct comparison. However, we observe that while our models forecast steep declines over the next 30 years, the principal variant predicts a gradual decrease with a slight resurgence for the youngest cohorts. This is consistent with the higher predicted ASFRs at key reproductive ages (Figure 8). We also note that our projections align most closely with the low fertility NPP variant.

6 Discussion and conclusion

This paper proposes methodology for obtaining parity-specific projections of fertility rates within a Bayesian set-up, allowing explicit and coherent uncertainty quantification. The core of our approach is a GAM initially fitted to individual-level data, into which we incorporate population-level data. We achieve this integration through a sophisticated marginalisation process, enabling examination of the results at three levels—unmarginalised, marginalised and aggregate—for

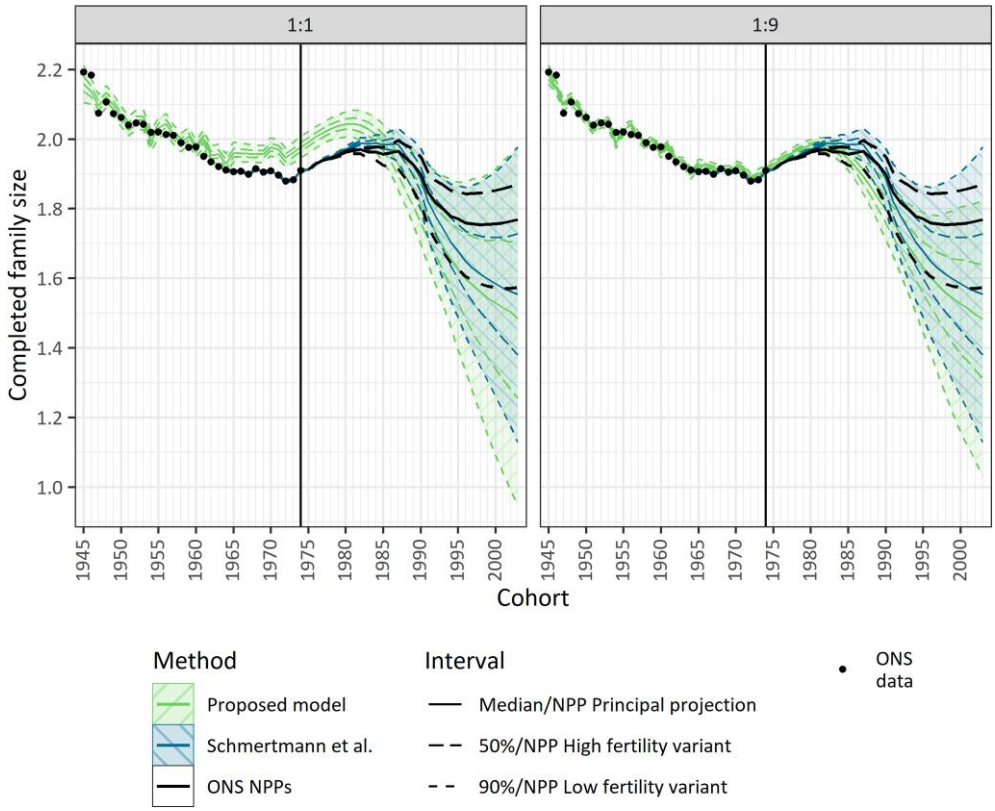

Figure 9. Completed family size posterior distributions under various integrated models and the model of Schmertmann et al. (2014a), as well as variants from the ONS 2018-based NPPs. Vertical lines indicate the last fully-observed cohort across the included datasets. ONS = Office for National Statistics; NPPs = National Population Projections.

differing relative contributions of the data sources. Our proposal advances the fertility projection literature in two main ways. First, we project fertility by parity, while current methods typically neglect to include such information. This allows us to mimic the sequential nature of childbearing decisions and capture the differences between the populations at risk of each birth order. Second, we incorporate individual-level data, selecting the variables to be included during the modelling process rather than beforehand, as has been done in the existing literature. This flexible and general approach is also principled, as the choice of variables is data-driven.

As, to the best of our knowledge, such a fertility projection model has not been previously proposed, our main discussion points concern the methodological challenges posed and how we address them. An important challenge is how best to obtain expressions of parity-specific birth probabilities by age and cohort when additional individual-level variables, here qualification and time since last birth, are involved. The aforementioned marginalisation approach is also used in the fertility estimation models of Handcock et al. (2005) and Rendall et al. (2009). These papers use the observed proportions to take a weighted average of the birth probabilities, but we found this to be unsatisfactory for two reasons. First, it does not translate well to a forecasting context, unless a crude approach such as freezing the proportions is applied. Second, if the proportions progress erratically, as we found for parity 3+ across age at small time since last birth values, this behaviour filters through to the marginalised level. We therefore develop multinomial logit projection models for the covariate probabilities that enforce smoothness across age (Sections 4.2 and 4.3). These models fit the observed counts closely, but their determination adds considerable complexity to our modelling process.

Following this marginalisation, another key challenge is how best to combine the individual- and population-level data sources. Due to a lack of knowledge as to the direction or amount of

bias in either dataset, it is not appropriate to use one source to constrain the inferences from the other, which tends to be done in previous work (Handcock et al., 2005; Rendall et al., 2009; Zhang & Bryant, 2019). We opt to take a logarithmic pooling approach applied to the likelihoods, governed by the critical pooling weight w (Section 4.4). Reweighting the likelihood allows us to control the balance of the information provided by both data sources in an intuitive way, with the value of w interpretable in relation to the relative importance of the datasets. However, the resulting combined likelihood, which is non-standard in that it is constructed from two data sources rather than a single source, can be viewed as somewhat abstract.

A further challenge is that of determining an appropriate value for w. We examine the sensitivity of projections to w, presenting marginalised forecasts for a range of values (Section 5.1); this is similar in spirit to Rendall et al. (2009), where various constraint priors are compared. This approach is manageable here because we only have one weight to vary; however, it would become impracticable were we to combine a larger number of data sources, and it also does not allow the weight(s) to be informed by the model itself (Carvalho et al., 2023). Various methods to learn about such weights more formally have been proposed in the literature, including cross-validation (Fletcher et al., 2019; Wang & Zidek, 2005) and modelling the weights explicitly as parameters, with associated priors in a Bayesian context (Carvalho et al., 2023). In our case, the need to use Monte Carlo methods to approximate the posterior, together with the highly contrasting sample sizes in the two data sources, mean that a cross-validation approach would likely prove problematic (Guo et al., 2012). A preliminary attempt to specify a prior for w in the parity 3+ model resulted in a posterior mean roughly corresponding to a 70:1 model (see Section 4.4), which would seem to give insufficient weight to the population-level data. Further work on methodology to estimate w is planned.

The final challenge is how best to additionally marginalise across parity, thus obtaining projections at the aggregate level that can be contrasted with existing methods. We achieve this by cumulating cohort fertility as in Smallwood (2002). This is not ideal as it assumes invariance of the parity distribution to mortality and migration in the population, which could have a significant impact on the results, particularly in the case of migration. Further information on the resulting caveats is provided in the methods protocol for the Human Fertility Database, where a similar approach is taken to construct fertility tables (Jasilioniene et al., 2015). Comparing our projections of CFS with those from the model of Schmertmann et al. (2014a), we find that the intervals overlap extensively (Figure 9). This is reassuring given that Bohk-Ewald et al. (2018) found this model to outperform existing methods in terms of uncertainty quantification. There are some small differences however, possible reasons for which we describe below.

One explanation for the discrepancies could be that the cumulative contributions across parity exacerbate the projected decline under our proposed model. Related to this, we note that close alignment with the population-level data at the aggregate level (i.e. the ASFRs) necessitates the same at the parity-specific level, i.e. by choosing a larger value of w. This is problematic because we have less confidence in the parity-specific rates compared to the ASFRs, and increasing w leads to overfitting and faster-changing projections (Section 5.1). Thus, future work should explore a joint approach, whereby the ASFRs could be constrained directly using the population-level estimates, and the CFS constrained to change in line with past trends (Figure 9) or expert opinion, for example from the NPP consultations (ONS, 2022a). This could help to inform about biases in the parity-specific rate estimates, increase forecast precision, and reduce forecast sensitivity to the amount of training data (Sections 5.3 and 5.4).

We make some additional suggestions for future extensions of this work. First, a key limitation is that we only present a single implementation of our proposed model—it is vital to investigate the robustness and generalisability of our methodology when applied to alternative data sources and to data from different countries. Of particular interest are countries where the required data are less accessible and may therefore be coarser or cover a shorter time period, for example less-developed countries. It is likely that the ability to constrain the ASFRs directly, an area of future work described above, will be essential if there is limited or even no availability of population-level parity-specific rates. Second, in this paper our geography of interest is England and Wales as a whole. However, there is also considerable subnational variation in fertility patterns and levels (e.g. see Campisi et al., 2020; Fiori et al., 2014). Accounting for such differences would rely on having sufficiently detailed data at the individual and/or population levels; it would also greatly

increase model complexity, which may not be justifiable. Depending on the research question, the geography and variables of interest might change. Further work could therefore include country, region and other variables in the model selection process so that they can be chosen if they substantially improve model fit.

In conclusion, combining individual- and population-level data in our Bayesian parity-specific fertility projection model generates plausible probabilistic forecasts for population subgroups determined by characteristics such as qualification and time since last birth, despite their absence in the population-level data. This has richness and value in itself, but also requires sophisticated methodology to perform the marginalisation. A comparison with one of the best-performing existing models indicates that our model predicts with reasonable accuracy and uncertainty; however, integrated modelling of the parities is needed to determine whether individual-level and/or parity information can improve predictive performance. Despite Bohk-Ewald et al. (2018) finding minimal accuracy gains in more complex cohort fertility forecasting methods, if a model incorporating such information can be fitted with reasonable efficiency and forecast with comparable reliability, the additional ability to provide plausible breakdowns by parity and/or individual-level variables would prove extremely valuable.

Acknowledgments

The authors thank two anonymous reviewers and an associate editor for their helpful comments on earlier versions of this paper. The UKHLS is an initiative funded by the Economic and Social Research Council (ESRC) and various government departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. We use code provided by Ann Berrington and Juliet Stone, which was used in the analyses for Berrington et al. (2015), to prepare the fertility histories. We acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. This paper extends work originally undertaken by the first author as part of their PhD, entitled 'Stochastic modelling and projection of age-specific fertility rates' (see funding information below). Earlier versions of this work have been presented at the British Society for Population Studies Annual Conference in September 2021 and the International Population Conference in December 2021.

Conflict of interest: None declared.

Funding

The doctoral programme of the first author was funded by the Engineering and Physical Sciences Research Council (award reference 1801045). This work was partly supported by the Economic and Social Research Council (ESRC) FertilityTrends project (grant ES/S009477/1), the ESRC Centre for Population Change (Transition funding) (grant ES/R009139/1), and the CPC-Connecting Generations Centre (grant ES/W002116/1).

Data availability

The individual-level data that support the findings of this study are available from the UK Data Service. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available at https://ukdataservice.ac.uk/ with the permission of the UK Data Service. The population-level data that support the findings of this study are openly available from the Office for National Statistics (https://www.ons.gov.uk/; specific references are provided in the paper). Code to replicate all results in this paper can be accessed at https://doi.org/10.5281/zenodo.8364164. Where code has been sourced from a third party, full details of where to obtain the code and any modifications made by the authors has been provided.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C.*

References

Abel G. J., Barakat B., KC S., & Lutz W. (2016). Meeting the Sustainable Development Goals leads to lower world population growth. *Proceedings of the National Academy of Sciences*, *113*(50), 14294–14299. https://doi.org/10.1073/pnas.1611386113

Balbo N., Billari F. C., & Mills M. (2013). Fertility in advanced societies: A review of research. *European Journal of Population*, *29*(1), 1–38. https://doi.org/10.1007/s10680-012-9277-y

Berg D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, *23*(2), 129–143. https://doi.org/10.1002/asmb.658

Berrington A., Stone J., & Beaujouan E. (2015). Educational differences in timing and quantum of childbearing in Britain: A study of cohorts born 1940–1969. *Demographic Research*, *33*, 733–764. https://doi.org/10.4054/DemRes.2015.33.26

Bijak J., & Bryant J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, *70*(1), 1–19. https://doi.org/10.1080/00324728.2015.1122826

Bohk-Ewald C., Li P., & Myrskylä M. (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences*, *115*(37), 9187–9192. https://doi.org/10.1073/pnas.1722364115

Booth H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, *22*(3), 547–581. https://doi.org/10.1016/j.ijforecast.2006.04.001

Campisi N., Kulu H., Mikolai J., Klüsener S., & Myrskylä M. (2020). Spatial variation in fertility across Europe: Patterns and determinants. *Population, Space and Place*, *26*(4), e2308. https://doi.org/10.1002/psp.2308

Carvalho L. M., Villela D. A. M., Coelho F. C., & Bastos L. S. (2023). Bayesian inference for the weights in logarithmic pooling. *Bayesian Analysis*, *18*(1), 223–251. https://doi.org/10.1214/22-BA1311

Chaudhuri S., Handcock M. S., & Rendall M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B*, *70*(2), 311–328. https://doi.org/10.1111/j.1467-9868.2007.00637.x

Currie I. D., Durban M., & Eilers P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, *4*(4), 279–298. https://doi.org/10.1191/1471082X04st080oa

de Beer J. (1985). A time series model for cohort data. *Journal of the American Statistical Association*, *80*(391), 525–530. https://doi.org/10.1080/01621459.1985.10478149

Eilers P. H. C., & Marx B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121. https://doi.org/10.1214/ss/1038425655

Ellison J. (2021). *Stochastic modelling and projection of age-specific fertility rates [PhD thesis]*. University of Southampton, School of Mathematical Sciences. https://eprints.soton.ac.uk/450468/.

Ellison J., Berrington A., Dodd E., & Forster J. J. (2022). Investigating the application of generalized additive models to discrete-time event history analysis for birth events. *Demographic Research*, *47*(22), 647–694. https://doi.org/10.4054/DemRes.2022.47.22

Ellison J., Dodd E., & Forster J. J. (2020). Forecasting of cohort fertility under a hierarchical Bayesian approach. *Journal of the Royal Statistical Society: Series A*, *183*(3), 829–856. https://doi.org/10.1111/rssa.12566

Ferrara G., & Vidoli F. (2017). Semiparametric stochastic frontier models: A generalized additive model approach. *European Journal of Operational Research*, *258*(2), 761–777. https://doi.org/10.1016/j.ejor.2016.09.008

Fiori F., Graham E., & Feng Z. (2014). Geographical variations in fertility and transition to second and third birth in Britain. *Advances in Life Course Research*, *21*, 149–167. https://doi.org/10.1016/j.alcr.2013.11.004

Fletcher R. J., Hefley T. J., Robertson E. P., Zuckerberg B., McCleery R. A., & Dorazio R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710. https://doi.org/10.1002/ecy.2710

Guo P., Wang X., & Wu Y. (2012). Data fusion using weighted likelihood. *European Journal of Pure and Applied Mathematics*, *5*(3), 333–356. https://www.ejpam.com/index.php/ejpam/article/view/1676

Handcock M. S., Huovilainen S. M., & Rendall M. S. (2000). Combining registration-system and survey data to estimate birth probabilities. *Demography*, *37*(2), 187–192. https://doi.org/10.2307/2648120

Handcock M. S., Rendall M. S., & Cheadle J. E. (2005). Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Methodology*, *35*, 291–334. https://doi.org/10.1111/j.0081-1750.2006.00170.x

Hu F., & Zidek J. V. (2002). The weighted likelihood. *The Canadian Journal of Statistics*, *30*(3), 347–371. https://doi.org/10.2307/3316141

Imbens G. W., & Lancaster T. (1994). Combining micro and macro data in microeconomic models. *The Review of Economic Studies*, *61*(4), 655–680. https://doi.org/10.2307/2297913

Institute for Social and Economic Research. (2022). *Variable search*. University of Essex. https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation

Jasilioniene A., Jdanov D. A., Sobotka T., Andreev E. M., Zeman K., Shkolnikov V. M., and with contributions from Goldstein J., Nash E. J., Philipov D., & Rodriguez G. (2015). *Methods Protocol for the Human Fertility Database*. https://www.humanfertility.org/File/GetDocumentFree/Docs/methods.pdf

Kravdal Ø. (2001). The high fertility of college educated women in Norway: An artefact of the separate modelling of each parity transition. *Demographic Research*, *5*, 187–216. https://doi.org/10.4054/DemRes.2001.5.6

Li N., & Wu Z. (2003). Forecasting cohort incomplete fertility: A method and an application. *Population Studies*, *57*(3), 303–320. https://doi.org/10.1080/0032472032000137826

Lohmann C., & Ohliger T. (2019). The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models. *Journal of Forecasting*, *38*(5), 375–389. https://doi.org/10.1002/for.2545

Myrskylä M., Kohler H. P., & Billari F. C. (2009). Advances in development reverse fertility declines. *Nature*, *460*(7256), 741–743. https://doi.org/10.1038/nature08230

Norman P., Rees P., & Wohland P. (2014). The use of a new indirect method to estimate ethnic-group fertility rates for subnational projections for England. *Population Studies*, *68*(1), 43–64. https://doi.org/10.1080/00324728.2013.810300

ONS. (2019). *Zipped population projections data files, Great Britain and England and Wales*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/datasets/z2zippedpopulationprojectionsdatafilesgbandenglandandwales

ONS. (2020). *Fertility rates by parity 1934 to 2018, England and Wales*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/adhocs/11482fertilityratesbyparity1934to2018englandandwales

ONS. (2022a). *National population projections quality and methodology information (QMI)*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/nationalpopulationprojectionsqmi

ONS. (2022b). *User guide to birth statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/methodologies/userguidetobirthstatistics

Pedersen E. J., Miller D. L., Simpson G. L., & Ross N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, *7*, e6876. https://doi.org/10.7717/peerj.6876

Poole D., & Raftery A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, *95*(452), 1244–1255. https://doi.org/10.1080/01621459.2000.10474324

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raftery A. E., Lewis S. M., Aghajanian A., & Kahn M. J. (1996). Event history modeling of world fertility survey data. *Mathematical Population Studies*, *6*(2), 129–153. https://doi.org/10.1080/08898489609525426

Rendall M. S., Admiraal R., DeRose A., DiGiulio P., Handcock M. S., & Racioppi F. (2008). Population constraints on pooled surveys in demographic hazard modelling. *Statistical Methods and Applications*, *17*(4), 519–539. https://doi.org/10.1007/s10260-008-0106-8

Rendall M. S., Handcock M. S., & Jonsson S. H. (2009). Bayesian estimation of hispanic fertility hazards from survey and population data. *Demography*, *46*(1), 65–83. https://doi.org/10.1353/dem.0.0041

Reyes Santías F., Cadarso-Suárez C., & Rodríguez-Álvarez M. X. (2011). Estimating hospital production functions through flexible regression models. *Mathematical and Computer Modelling*, *54*(7–8), 1760–1764. https://doi.org/10.1016/j.mcm.2010.11.087

Schmertmann C., Zagheni E., Goldstein J. R., & Myrskylä M. (2014a). Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association*, *109*(506), 500–513. https://doi.org/10.1080/01621459.2014.881738

Schmertmann C., Zagheni E., Goldstein J. R., & Myrskylä M. (2014b). *Bayesian forecasting of cohort fertility*. http://schmert.net/cohort-fertility/

Smallwood S. (2002). New estimates of trends in births by birth order in England and Wales. *Population Trends*, *108*, 32–48. https://webarchive.nationalarchives.gov.uk/ukgwa/20160110122826/http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no–108–summer-2002/index.html

Stan Development Team. (2019). *RStan: The R interface to Stan* (R package version 2.19.2). http://mc-stan.org/.

Steele F. (2005). *Event history analysis*. National Centre for Research Methods Review Paper NCRM/004. http://eprints.ncrm.ac.uk/88/

Umlauf N., Klein N., & Zeileis A. (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, *27*(3), 612–627. https://doi.org/10.1080/10618600.2017.1407325

United Nations Development Programme. (2018). *Human development indices and indicators: 2018 statistical update—Technical notes*. https://hdr.undp.org/sites/default/files/data/2020/hdr2018˙technical˙notes.pdf

United Nations Development Programme. (2019). *Human Development Report 2019*. https://hdr.undp.org/content/human-development-report-2019

United Nations Development Programme, Department of Economic and Social Affairs, Population Division. (2019a). *Population facts No. 2019/5, December 2019: Potential impact of later childbearing on future population*. https://www.un.org/en/development/desa/population/publications/pdf/popfacts/PopFacts_2019-5.pdf

United Nations Development Programme, Department of Economic and Social Affairs, Population Division. (2019b). *Population facts No. 2019/6, December 2019: How certain are the United Nations global population projections?*. https://www.un.org/en/development/desa/population/publications/pdf/popfacts/PopFacts_2019-6.pdf

University of Essex Institute for Social and Economic Research, NatCen Social Research, Kantar Public. (2017). *Understanding Society: Waves 1–7, 2009–2016: Special Licence Access* [data collection] (7th ed.). UK Data Service. http://doi.org/10.5255/UKDA-SN-6931-7, Data downloaded July 19, 2018.

Vollset S. E., Goren E., Yuan C. W., Cao J., Smith A. E., Hsiao T., Bisignano C., Azhar G. S., Castro E., & Chalek J. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: A forecasting analysis for the Global Burden of Disease Study. *The Lancet*, *396*(10258), 1285–1306. https://doi.org/10.1016/S0140-6736

Waller L., Berrington A., & Raymer J. (2014). New insights into the fertility patterns of recent Polish migrants in the United Kingdom. *Journal of Population Research*, *31*(2), 131–150. https://doi.org/10.1007/s12546-014-9125-5

Wang X., & Zidek J. V. (2005). Selecting likelihood weights by cross-validation. *The Annals of Statistics*, *33*(2), 463–500. https://doi.org/10.1214/009053604000001309

Wood S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC Press.

Zhang J. L., & Bryant J. (2019). Combining multiple imperfect data sources for small area estimation: A Bayesian model of provincial fertility rates in Cambodia. *Statistical Theory and Related Fields*, *3*(2), 178–185. https://doi.org/10.1080/24754269.2019.1658062