Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEACH ARTICLE

Penalised smoothing splines resolve the curvature identifiability problem in age-period-cohort models with unequal intervals.

Department of Mathematical Sciences,
University of Bath, Bath, UK

Correspondence
Connor Gascoigne, Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK. Email: c.gascoigne@bath.ac.uk

Summary

Age-period-cohort (APC) models are frequently used in a variety of health and demographic related outcomes. Fitting and interpreting APC models to data in equal intervals (equal age and period widths) is non-trivial due to the structural link between the three temporal effects (given two, the third can always be found) causing the well-known identification problem. The usual method for resolving the structural link identification problem is to base a model on identifiable quantities. It is common to find health and demographic data in unequal intervals, this creates further identification problems on top of the structural link. We highlight the new issues by showing that curvatures which were identifiable for equal intervals are no longer identifiable for unequal data. Furthermore, through extensive simulation studies, we show how previous methods for unequal APC models are not always appropriate due to their sensitivity to the choice of functions used to approximate the true temporal functions. We propose a new method for modelling unequal APC data using penalised smoothing splines. Our proposal effectively resolves the curvature identification issue that arises and is robust to the choice of the approximating function. To demonstrate the effectiveness of our proposal, we conclude with an application to UK all-cause mortality data from the Human mortality database (HMD).

KEYWORDS:
Age-period-cohort models, identifiability, penalised smoothing splines, unequal intervals

1 | INTRODUCTION

Age-period-cohort (APC) models are used to interpret the effects of the most influential temporal trends on incidence and mortality rates for a multitude of diseases. Age effects are a measure of attrition on one's body as they get older, period (time of the event) effects reflect short term exposures (e.g., new treatments) and a (commonly birth) cohort effect is a long-term exposure (e.g., smoking views). We use obesity as an example of how all three temporal effects relate to a major health concern. Obesity is a measure of an individual's body mass index (BMI) with those greater than or equal to 30 being classified as obese. The most recent health survey from the UKs National Health Service (NHS) found 68% and 60% of adult men and women are classified as obese, respectfully.[1] Weight of an individual increases with age, and in recent years there has been an increasing trend of obesity. These together make for a cohort effect where those born in more recent cohorts have an increased risk to being obese and getting there at an earlier age.

APC models are affected by an identifiability problem due to the linear dependence between the three temporal terms. For example, a birth cohort can always be found by subtracting the age of the individual from the year the response was taken. The result of this linear dependence is that the three linear trends are impossible to disentangle from one another without the use of additional information. We will call this problem the "structural link identification problem" (or structural link for short).

Commonly, APC models are considered when data comes tabulated in equal widths (equal intervals). Appropriate solutions to the structural link are based on reparameterising the APC modelling into estimable quantities. When time is considered discrete (most common), each temporal term is modelled as a factor with levels for each time interval; Holford pioneered a solution based on estimable curvatures[2] (terms that are orthogonal to a linear term) for each temporal effect, whilst Kuang, Nielson and Nielson based a solution on estimable second differences[3] (a discrete version of the second derivative). When time is considered continuous, the temporal terms are often modelled by approximate smooth functions. Carstensen defined a set of estimable quantities, like Holford's curvatures, to fit APC models.[4] Smith and Wakefield offer a comprehensive review on APC models for data aggregated in equal intervals.[5]

Less commonly, APC models are considered when data comes tabulated in unequal intervals. This contrasts the fact that many providers of health and demographic data frequently release data tabulated in unequal intervals. For example, the UK's office of national statistics (ONS) releases all-cause mortality data in single-year age and period[6] and, for a finer understanding of seasonality, weekly periods, and five-year age groups.[7] In addition, the Demographic and Health Surveys (DHS) release data for monthly ages and yearly periods.[8] APC models fit to unequal data are less common as the model fitting process induces more identification issues (on top of the structural link) that are displayed by a cyclic pattern in the previously estimable functions.[9] Figure 1 shows the cohort curvature estimates when a factor model is fit to simulated unequal data. Note the cyclic pattern, this could be due to the underlying phenomena of interest being modelled. However, later in the paper it is shown the cyclic pattern is more likely caused by the further identification problems present when modelling data aggregated into unequal data.

As Figure 1 alludes to, factor-based approaches based on estimable quantities that worked for data in equal intervals are no longer appropriate when data comes in unequal intervals. A proposed method to model APC data in unequal intervals is to model the temporal terms with approximate smooth functions such as smoothing splines.[9,4] These may resolve the issue but raise additional questions about how to specify the smooth functions and if they are sensitive to the choice of specification. The use of a penalised spline has been recognised as potentially preferable solution to the cyclic pattern than just a smoothing spline alone,[10] but has not been fully explored until now. Another approach is to collapse unequal intervals into equal intervals but in many cases, this causes a large amount of information lost which decreases the reliability of the results.

The purpose of this paper is to propose a method to modelling APC data that comes in unequal intervals. The method we propose addresses all identification problems present for unequal data, maintains clarity on what is and is not estimable, has a clear interpretation, and is robust to the choice of function used to approximate each temporal term. We propose approximating each temporal term as a continuous function, reparameterise each into a linear and orthogonal curvature and when modelling the curvatures, include a penalty on the second derivative (a measure of "wiggliness") of the estimate. Using continuous functions in a reparameterised APC model is not new. For example, Heuer performed a simulation study for APC models fit to data in unequal intervals using continuous functions to model period and cohort curvature terms,[11] and Carstensen promotes the use of continuous functions in his reparameterisation.[4] However, the novelty we are proposing relates to the use, specification, and implementation of a penalty on estimable terms within a reparameterised APC model.

We show how to correctly construct a penalty that is only penalising the curvature terms after the reparameterisation and explain how to implement it practically. Via simulation studies, we confirm the use of a penalty in a reparameterised APC model is appropriate for fitting models to data both equally and unequally aggregated. A sensitivity analysis is used to demonstrate that the inclusion of a penalty provides robustness to how the continuous functions are specified. The same robustness is not present in the absence of a penalty in the sensitivity analysis highlighting the necessity of the penalty function when considering data unequally aggregated in an APC model.

The remainder of the article is organised as follows. In Section 2, we review the identification problem for data aggregated in equal intervals and introduce our new reparameterisation scheme. Section 3 is a comprehensive simulation study for the case when data comes in equal intervals. Section 4, we review the curvature identification problem that arises from unequal intervals and show through theoretical and simulation results how the proposed method relieves this added identification problem. Finally, we conclude with an application to all-cause mortality data in the UK in Section 5 and a conclusion in Section 6.

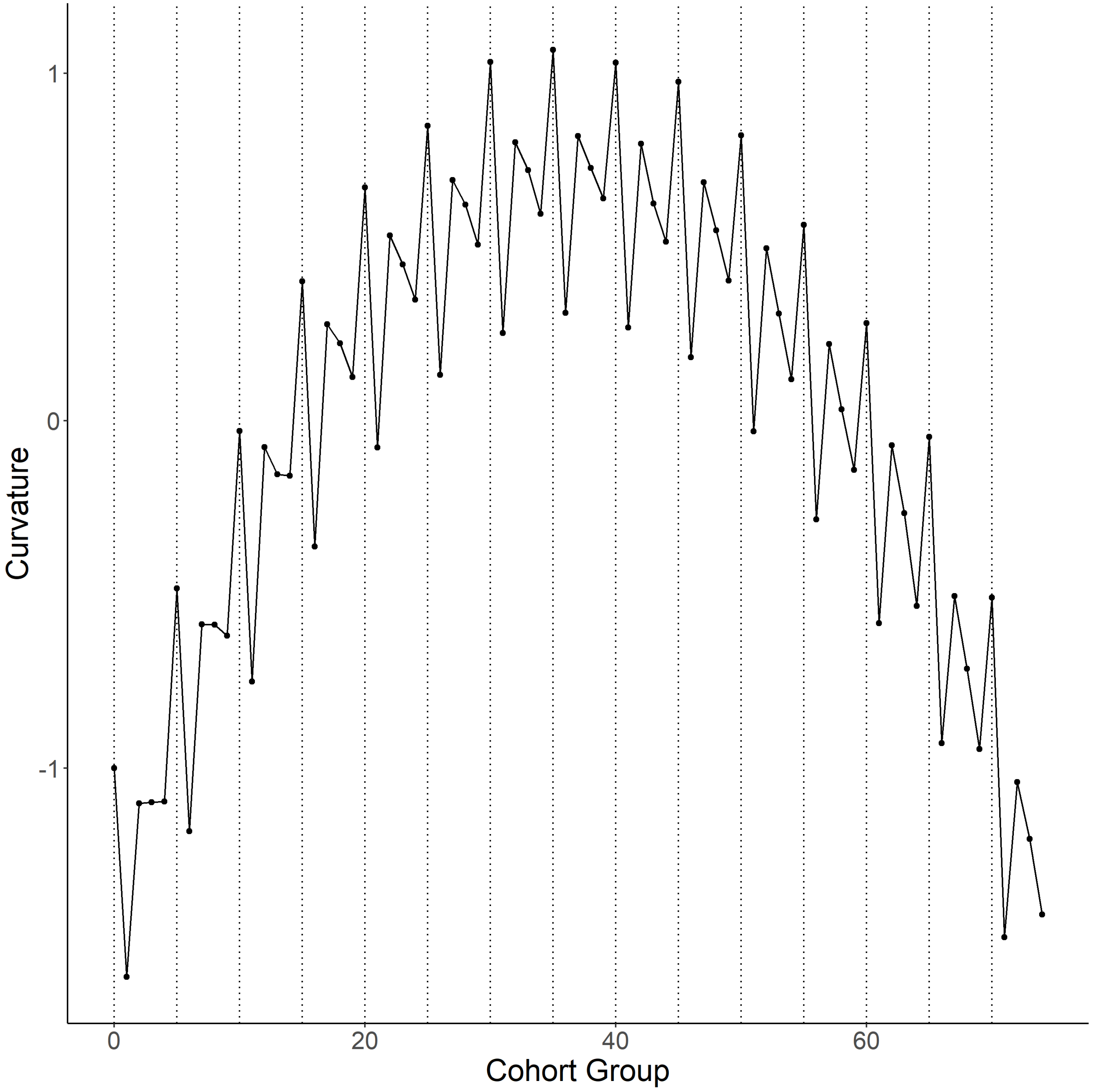


FIGURE 1 Cohort curvature estimate from fitting an APC model reparameterised into linear terms and their orthogonal curvatures to simulated unequal interval APC data.

2 | METHOD

2.1 | Identification Problems

We begin by discussing an APC model for data equally aggregated, referred to as 'equal intervals'. There are two types of identification problems in this model. The first is well-known and due to including an intercept along with more than one smooth function (or factor) in a model. The second and more serious is due to the structural link. The structural link occurs since given any two of age, period, or cohort, the third can always calculated. Commonly, birth cohort is found by taking the difference between year of event and age, $c = p - M \times a$ where M is the ratio of age interval to period interval. For equal intervals $M = 1$, this simplifies to $c = p - a$.

Table 1 shows how cohort index varies when age and period are aggregated into equal intervals. With age increasing from bottom to top and period left to right, a cohorts progression can be traced on the bottom left to top right diagonal. The earliest cohort is top left (oldest age with the first year) and the most recent cohort is bottom right (youngest age with most recent year).

8	1	2	3	4	5	6	7	8
7	2	3	4	5	6	7	8	9
6	3	4	5	6	7	8	9	10
5	4	5	6	7	8	9	10	11
4	5	6	7	8	9	10	11	12
3	6	7	8	9	10	11	12	13
2	7	8	9	10	11	12	13	12
1	8	9	10	11	12	13	14	15
Age	1	2	3	4	5	6	7	8
	Period							

TABLE 1 Cohort indexing for age-period data table where age is grouped $M = 1$ times larger than period. The cohort index is defined using $c = M \times (A - a) + p$ where $A = 8$ to fix the first cohort to be 1.

Let y_{ap} be response from age group a and period group p where $a = 1, 2, \dots, A$ and $p = 1, 2, \dots, P$. A cohort index is not explicitly defined due to the structural link, but is calculated $c = 1, 2, \dots, C = M \times (A - a) + p$. A continuous APC model is

$$g\left(\mu_{ap}\right) = f_A(a) + f_P(p) + f_C(c) \tag{1}$$

where $g(\cdot)$ is the link function, $\mu_{ap} \equiv \mathbb{E}\left[y_{ap}\right]$ is equivalent to the expected value of the response and f_A, f_P and f_C are the smooth functions of age a, period p and cohort c. The APC identification problem means we can add a constant and linear trend to each function without affecting the overall linear predictor. Consider the following functions[4]

$$\begin{aligned}
\tilde{f}_A(a) &= f_A(a) - \text{const}_1 + \text{const}_3 M a \\
\tilde{f}_P(p) &= f_P(p) + \text{const}_1 + \text{const}_2 - \text{const}_3 p \\
\tilde{f}_C(c) &= f_C(c) - \text{const}_2 + \text{const}_3 c
\end{aligned}$$

where const_1 and const_2 are due to the identification for including more than one smooth function and const_3 is due to the structural link. The overall linear predictor is invariant to the inclusion of these constants since

$$\begin{aligned}
g\left(\tilde{\mu}_{ap}\right) &= \tilde{f}_A(a) + \tilde{f}_P(p) + \tilde{f}_C(c) \\
&= \left[f_A(a) - \text{const}_1 + \text{const}_3 M a\right] + \left[f_P(p) + \text{const}_1 + \text{const}_2 - \text{const}_3 p\right] + \left[f_C(c) - \text{const}_2 + \text{const}_3 c\right] \\
&= \left[f_A(a) + \text{const}_3 M a\right] + \left[f_P(p) - \text{const}_3 p\right] + \left[f_C(c) + \text{const}_3 c\right] \\
&= f_A(a) + f_P(p) + f_C(c) = g\left(\mu_{ap}\right)
\end{aligned}$$

where $c = p - M \times a$ is used in the fourth line.

Many reparameterisation schemes are based off of an identifiable set of quantities. The first derivatives are not identifiable[12,13] but the second derivatives, or more generally curvatures, are identifiable. The age curvature is expressed

$$\widetilde{f_{A_C}}(a) \equiv \tilde{f}''_A(a) = f''_A(a) \equiv f_{A_C}(a)$$

where the subscript "C" denotes the curvature. The terms for period and cohort curvature are analogous.

2.2 | Univariate temporal model

For the purpose of development, we shall first focus on a single temporal function for age. A univariate temporal model for age can be expressed as

$$g\left(\mu_a\right) = f_A(a) \tag{2}$$

for f_A, a smooth function of covariate a for age.

A popular set of functions used to approximate the smooth functions are splines: sums of polynomial functions called basis functions, which are based on a selection of points called knots. Within APC modelling, the `Epi`[4] package in `R`[14] fits several splines bases to continuous functions of APC models without penalisation. Carstensen also incorporated his methods into a package in `Stata` with extensions to include covariates.[15]

To approximate f_A in Eq.(2), the user specifies basis functions, and the model fitting process produces estimates for the weights of said basis functions. Given the basis $b_i(a)$, the i^{th} basis function, f_A is approximated with a spline as follows

$$\hat{f}_A(a) = \sum_{i=1}^{I} b_i(a)\hat{\beta}_i$$

where $\boldsymbol{I}$ is the number of basis functions and $\hat{\beta}_i$ is the estimate of the unknown weights.

Estimates of the true function can be found using a penalised iterative re-weighted least squares (PIRLS) algorithm to produce $\hat{\beta}_i$.[16] PIRLS is used to find an estimate $\hat{\boldsymbol{f}}_A$ that minimises the objective function

$$\boldsymbol{D}\left(f_A(a)\right) + \lambda_A \int f_A''(a)^2\, da$$

where $\boldsymbol{D}\left(f_A(a)\right)$ is the deviance (square of the difference between the saturated log-likelihood and model log-likelihood) of the model and $\lambda_A \int f_A''(a)^2\, da$ is a penalty term on the second derivative "wiggliness" of f_A with smoothing parameter λ_A. For more details, see Chapter 4 Wood (2017).[16] By representing the smooth function via a spline basis, the smooth itself can be written

$$f_A(a) = \sum_{i=1}^{I} b_i(a)\beta_i = \boldsymbol{X\beta}$$

for $\boldsymbol{X}$ an $n \times \boldsymbol{I}$ matrix and $\boldsymbol{\beta}$ an $\boldsymbol{I} \times 1$ vector of parameters. The penalty function can be expressed as

$$\int f_A''(a)^2\, da = \boldsymbol{\beta}^T \int \boldsymbol{b}^T(a)\,\boldsymbol{b}(a)\, da\boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{S}_A \boldsymbol{\beta}$$

where $\boldsymbol{S}_A = \int \boldsymbol{b}^T(a)\,\boldsymbol{b}(a)\, da$ is the penalty matrix.

Penalising estimates of the smooth function reduces the effect of over fitting (e.g., from choosing too many bases to represent the smooth function) as over-fit functions are often "wigglier" than those under-fit and hence penalised greater. The smoothing parameter controls the trade-off between smoothness of the estimated smooth and closeness to the data. If $\lambda_A = 0$, there is no cost for fitting complicated functions while $\lambda_A \to \infty$ gives the maximum cost for fitting a complicated function, and $\hat{f}_A$ is a straight line.

2.3 | Orthogonalization

Often an intercept is included alongside smoothers; this causes identifiability problems that can be resolved via reparameterisation. A 'sum-to-zero' constraint orthogonalizes the smooth to an intercept term such that $\boldsymbol{1}^T\boldsymbol{X\beta} = \boldsymbol{0}$, avoiding any intercept related identification problems. The constraint is applied by constructing an $\boldsymbol{I}\times(\boldsymbol{I}-1)$ matrix $\boldsymbol{Z}$ through the QR-decomposition of $\left(\boldsymbol{1}^T\boldsymbol{X}\right)^T$. The smooth is reparameterised by using $\boldsymbol{XZ}$ and $\boldsymbol{Z}^T\boldsymbol{SZ}$ as its model and penalty matrices; for more details, see Chapter 5 Wood (2017).[16]

The parameter space of f_A can be split further into a linear slope and parameters corresponding to orthogonal curvatures.[2] In the following, orthogonality is defined with respect to the usual inner product $\langle x|y\rangle = \sum_i x_i y_i$; see Carstensen for a discussion on the choice of inner product used in the orthongonalization.[4] In the same vein as the intercept reparameterisation, define a $2\times\boldsymbol{I}$ array consisting of a constant and vector of all ages for the intercept and linear terms, $[\boldsymbol{1} : \boldsymbol{a}]$. Consequently, a $(\boldsymbol{I}-1)\times(\boldsymbol{I}-2)$ matrix $\boldsymbol{Z}$ is calculated by the QR-decomposition of $\left([\boldsymbol{1} : \boldsymbol{a}]^T\boldsymbol{X}\right)^T$, and the smooth f_A is reparameterised using $\boldsymbol{A}_C = \boldsymbol{XZ}$ and $\boldsymbol{S}_{A_C} = \boldsymbol{Z}^T\boldsymbol{SZ}$ as its model and penalty matrices.

After the intercept and linear slope reparameterisation, the form of the age-model is

$$g\left(\mu_a\right) = \beta_0 + a\beta_{A_L} + f_{A_C}(a)$$

where β_0 and β_{A_L} are the parameters for the intercept and slope and f_{A_C} is the smooth of covariate a orthogonal to the intercept and linear term. In matrix form,

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} = \left[\mathbf{1} \,:\, \boldsymbol{a} \,:\, \boldsymbol{A}_C\right] \begin{bmatrix} \beta_0 \\ \beta_{A_L} \\ \boldsymbol{\beta}^T_{A_C} \end{bmatrix}$$

where β_0, β_{A_L} and $\boldsymbol{\beta}_{A_C}$ are the parameters for the intercept, slope and curvature terms defined by the partitions $\mathbf{1}$, $\boldsymbol{a}$ and $\boldsymbol{A}_C$ of the model matrix. The smooth function f_{A_C} has the associated penalty $\lambda_A \boldsymbol{\beta}^T_{A_C} \boldsymbol{S}_{A_C} \boldsymbol{\beta}_{A_C}$.

Figure 2 shows how a spline basis for $a = 1, \ldots, 20$ with $i = 1, \ldots, 5$ basis functions changes after each reparameterisation. The first two rows, `b0` and `b1`, show the basis functions that capture the constant and linear behaviour of the spline, respectfully, and the remaining rows, `b2`, `b3` and `b4`, capture the higher order behaviour of the spline. The first column are the bases before any orthogonalization and the second and third columns show the bases after being orthogonalized to a constant and a constant and a linear term, respectfully. More details on specification of the spline basis will follow at the end of this section. When orthogonalizing the basis with respect to a constant and a linear trend, any part of the basis that captures these trends are removed. This is shown in Figure 2 by `b0` being removed after orthogonalization to an intercept and both `b0` and `b1` being removed after orthogonalization to both an intercept and linear trend. The bases that capture the higher order behaviour are also altered (e.g., rotated) since they may capture some form of a constant and linear trend.

2.4 | Age-period-cohort modelling

After reparameterising period and cohort in the same manner as age, an estimable APC model is written,

$$g\left(\mu_{ap}\right) = \beta_0 + s_1\beta_1 + s_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c) \tag{3}$$

where s_1 and s_2 are two of the three temporal slopes with parameters β_1 and β_2, and $f_{A_C}(a)$, $f_{P_C}(p)$ and $f_{C_C}(c)$ are the smooths for the age, period, and cohort curvatures. If all three slopes are included in the above reparameterisation, the model is over-parameterised. By dropping any one of the three slopes, the model is no longer over-parameterised, and the scheme is based off the identifiable curvatures that are invariant to the choice of slope dropped.[2] That is, dropping any of the age, period or cohort slopes does not change the estimates of the curvatures.

To generalise what parameters are estimable, let s_a, s_p and s_c be the respective age, period, and cohort linear terms. Any linear combination of $\kappa_1 s_a + \kappa_2 s_p + \left(\kappa_2 - \kappa_1\right) s_c$ is estimable for arbitrary κ_1 and κ_2.[2] While individual slopes cannot be estimated, the recommendation from Holford is to drop one slope as the effect of the dropped slope is included in the remaining two, which is *ad-hoc*.

Reparameterisations using curvatures orthogonal to linear slopes,[2] second differences[3] (second differences are the discrete way to define local curvature, like the continuous second derivatives) and period and cohort terms orthogonal to linear trends[4] provide systematic solutions to the APC problem. In each scheme, an arbitrary choice is made: which linear slope to drop,[2] which three baseline rates to choose[3] and which term and reference term to use.[4] Consequently, we refer to the aforementioned schemes as 'overall non-arbitrary' - they are based on a set of identifiable quantities but require an arbitrary choice during the reparameterisation process.

Depending on which of the two slopes are kept in, the interpretation of the model changes. Commonly the age slope is often retained due to age's importance in most health concerns. If the cohort slope is dropped, the APC model is "cross-sectional", i.e.,

$$g\left(\mu_{ap}\right) = \beta_0 + a\beta_1 + p\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c)$$

and if the period slope is dropped, the APC model is "longitudinal", i.e.,

$$g\left(\mu_{ap}\right) = \beta_0 + a\beta_1 + c\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c).$$

In the remainder of the paper, we do not concern ourselves with the interpretation of the model based off of the slope dropped and choose to drop the cohort slope in all subsequent models for consistency.

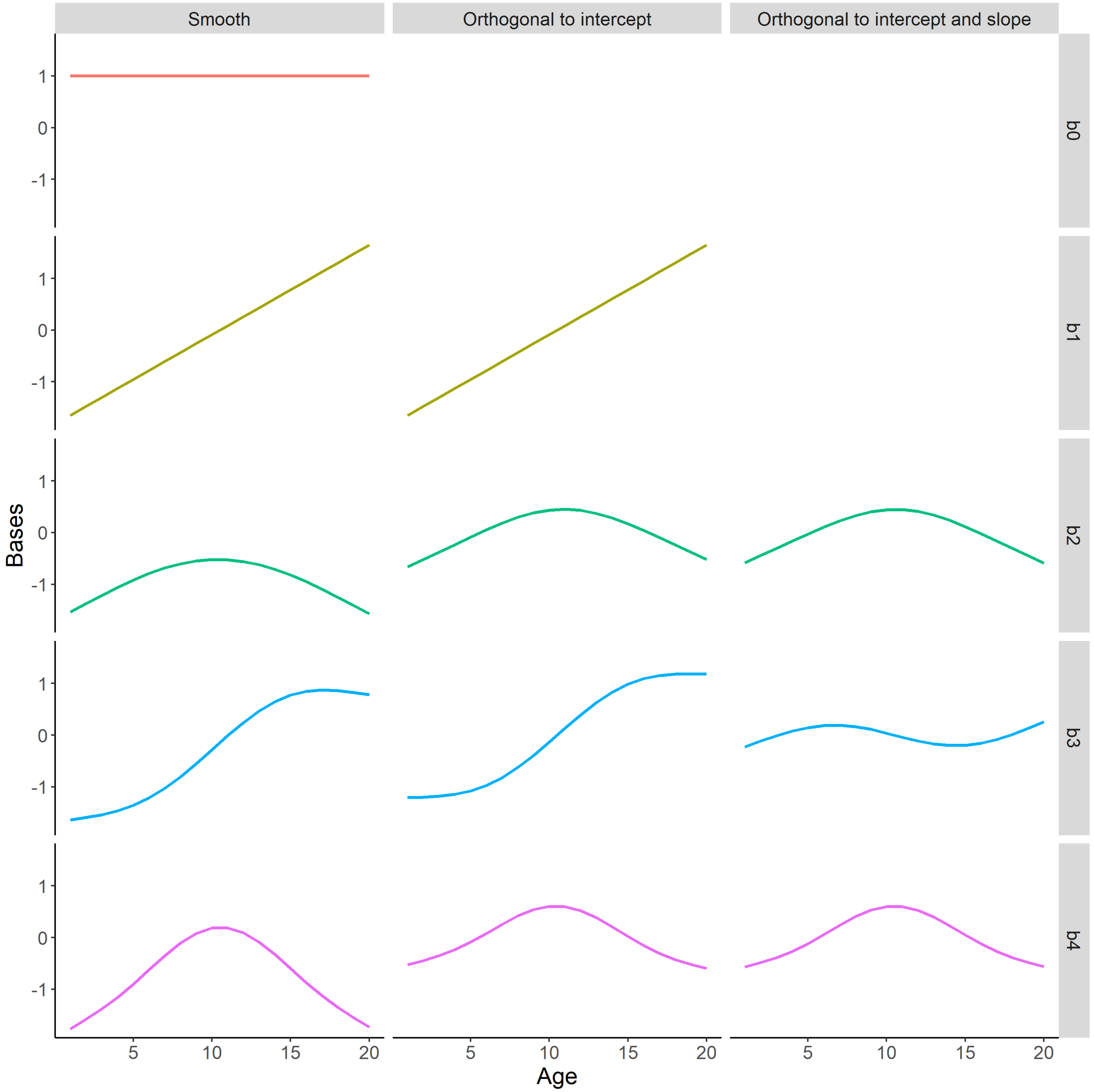


FIGURE 2 Thin plate regression spline basis before any reparameterisation, after an intercept reparameterisation and after an intercept and slope reparameterisation.

For models with multiple smooth functions, the penalty in the objective function is the addition of each individual smooths penalty function. For APC models reparameterised as above, the penalty function is

$$\lambda_a \int_a f''_{A_C}(a)^2\, da + \lambda_p \int_p f''_{P_C}(p)^2\, dp + \lambda_c \int_c f''_{C_C}(c)^2\, dc,$$

or in matrix form,

$$\lambda_a \boldsymbol{\beta}_{A_C} \boldsymbol{S}_{A_C} \boldsymbol{\beta}_{A_C} + \lambda_p \boldsymbol{\beta}_{C_C} \boldsymbol{S}_{P_C} \boldsymbol{\beta}_{P_C} + \lambda_c \boldsymbol{\beta}_{P_C} \boldsymbol{S}_{C_C} \boldsymbol{\beta}_{C_C}.$$

2.5 | Implementation

There are many types of spline basis functions one might use with common choices including thin plate regression splines and cubic regression splines. Thin plate regression splines smooth with respect to any number of covariates and do not need the knots to be specified *a priori*; however, thin plate regression splines are computationally costly and are not invariant to rescaling of the covariate. Cubic regression splines are computationally cheap with directly interpretable parameters but can only model one covariate at a time and require the knots to be predefined. For more details on these bases and examples of others, see Chapter 5 Wood (2017).[16] In Figure 2, we use a thin plate regression spline as they clearly illustrate the bases that capture the constant and linear behaviour of the spline. Going forward, we use a cubic regression spline basis for the implementation in the remainder of the paper. We note that our proposed method is not basis specific and will work for other basis choices.

The APC model in Eq.(3) has a parametric component, the two included slopes, and a non-parametric component, the smooth functions of curvatures. Therefore, it can fit into the wider framework of a generalised additive model (GAM).[17] We implement the GAMs in the `mgcv` package[16] in `R`[14] which offers a wide range of spline bases to represent smooth functions and their penalties. An example formula for a GAM in `mgcv` with one parameteric component (`x1`) and two non-parameteric components (`x2` and `x3`) is `y ~ x1 + s(x2, bs=bs, k=x2k) + s(x3, bs=bs, k=x3k)`. Here `s()` is the call to a univariate smooth, `bs` is the argument to specify the basis to use (e.g. "`bs = "cr"`" for a cubic regression spline) and `k` is the basis dimensions (the knots in the case of a spline). To manually modify the model and penalty matrices, we utilise the `fit = FALSE` argument in `gam()`. This argument returns the full model, including the model and penalty matrices, before the model fitting process. We take model and penalty matrices returned here, modify them as required, and then perform the model fitting process. Code to replicate the following simulation studies and application analysis can be found at https://github.com/connorgascoigne/Unequal-Interval-APC-Models.

3 | SIMULATION STUDY

We now present a simulation study to demonstrate that the proposed model provides a suitable solution to the APC identification problem in the simplest, $M = 1$, scenario.

3.1 | Data

The simulation study is motivated by obesity rates and how they increase with age and in more recent years[18,19,20] as well as having an hereditary effect.[21,22] Obesity data is a common application of APC models as a range of different responses can be used. For example, a linear regression model can be used on weights.[23] Alternatively, body mass index (BMI) has been modelled via a linear regression model (using log-BMI) and a logistic regression model (indicator of a given BMI).[24] In addition, a Poisson model for counts of rare events can be used.

The shapes for the age, period and cohort effects are adopted from a simulation study for Gaussian data from Luo and Hodges.[23] We extend this study to include responses from binomial and Poisson paradigms. Specific choices for the simulation set up and distribution parameters in the binomial and Poisson cases are motivated by the UKs yearly obesity survey from the NHS.[1] The survey is of approximately 8000 adults grouped from 16-24 up to 75+.

Data is simulated for individuals in single-year age-period format. We consider time to be continuous and use the yearly midpoint when modelling. We define single-year ages between $[0, 60]$ and single-year period between $[0, 20]$ with a (relative to the period) cohort calculated using $c = p - a$. For the normal and Binomial distributions, N_{ap} reflects the number of individuals included in the survey which for each of the 60 ages is fixed to be 150. For the Poisson distribution, N_{ap} is typically the population at risk, but for consistency this will be kept at 150 as well.

The true functions of age, period, and cohort (identical to Luo and Hodges) to generate the Gaussian data are

$$\begin{aligned} h_A(a) &= 0.3a - 0.01a^2 \\ h_P(p) &= -0.04p + 0.02p^2 \\ h_C(c) &= 0.35c - 0.0015c^2. \end{aligned}$$

In order to use the same set of functions (so the curve shapes are consistent across distributions), the simulations for the binomial and Poisson case are altered via an offset and scaling factor

$$\text{offset} + \text{scale} \times \left[h_A(a) + h_P(p) + h_C(c)\right]$$

to match the obesity survey data. The expected responses for the binomial (overweight, BMI ≥ 25) and Poisson (obese, BMI ≥ 30) data reflect an average of approximately 64% and 28% of the UKs adult population, respectively. Furthermore, both sets of responses have approximately 20% difference between the age group with the smallest largest counts.

The data from each distribution is generated from

$$\begin{aligned} y^k_{nap} &\sim \text{Normal}\left(\mu_{ap}, 1\right) \\ y^k_{ap} &\sim \text{Binomial}\left(N_{ap}, \pi_{ap}\right) \\ y^k_{ap} &\sim \text{Poisson}\left(\lambda_{ap}\right) \end{aligned}$$

where $\mu_{ap} = 0 + \left[h_A(a) + h_P(p) + h_C(c)\right]/1$ for $n = \{1, \ldots, N_{ap} = 150\}$ and $k = \{1, \ldots, K = 100\}$ for each simulation. For the binomial and Poisson distributions, $\pi_{ap} = \text{expit}\left(0.4 + \left[h_A(a) + h_P(p) + h_C(c)\right]/50\right)$ and $\lambda_{ap} = N_{ap}\exp\left(-1.5 + \left[h_A(a) + h_P(p) + h_C(c)\right]/50\right)$, respectively. The range of binomial and Poisson responses are approximately 45% to 81% and 9% to 51%, respectively, which match the target percentages with ±20%.

In this paper we will only report on the results for binomial generated data. The results for the other distributions are in the Supplementary Material. Furthermore, the Supplementary Material contains an example of data generated without all three temporal trends present (cohort is missing). This example highlights that the issues are due to the structural link within the data rather than the re-parameterisation we propose.

3.2 | Models

To each of the S data sets, we fit the following models:

1. Factor (FA) Model: A factor version of an APC model is written $g\left(\mu_{ap}\right) = \beta_0 + \alpha_a + \tau_p + \gamma_c$ where β_0 is the overall level, α_a, τ_p and γ_c are the a, p and c levels of the age, period, and cohort factors, respectively. The interpretation of these factors if there was no structural link identification would be relative risks (for example, for age it is the difference between the overall mean and the a^{th} age group). Due to the structural link, the factors are unidentifiable and cannot be interpreted as such; consequently, this model was originally reparameterised into a set of linear trends and their orthogonal curvatures.[2] The factor version of the reparameterised APC model is,

$$g\left(\mu_{ap}\right) = \beta_0 + s_1\beta_1 + s_2\beta_2 + \alpha_{C_a} + \tau_{C_p} + \gamma_{C_p}$$

where s_1 and s_2 are the two chosen linear trends with slopes β_1 and β_2 and $\boldsymbol{\alpha}_C$, $\boldsymbol{\tau}_C$ and $\boldsymbol{\gamma}_C$ are the factor curvature terms. This original reparameterisation is used as a benchmark for comparison in the simulation study and is still widely used with summaries available from a user-friendly web tool https://analysistools.cancer.gov/apc/ from the National Cancer Institute.[25]

2. Smoothing spline models: Detailed in Section 2, a reparameterisation in the style of the FA model but on a continuous version of the APC model using smoothing splines on the curvatures

$$g\left(\mu_{ap}\right) = \beta_0 + s_1\beta_1 + s_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c)$$

where s_1 and s_2 are the two chosen linear trends with slopes β_1 and β_2, and $f_{A_C}(a)$, $f_{P_C}(p)$ and $f_{C_C}(c)$ are the smooth functions of curvature. The smooth functions are represented by cubic regression splines with the number of knots approximately 25% of the number of unique data points for each temporal effect, spaced at even intervals.

   (a) Regression Smoothing spline (RSS): This is the smoothing spline model fit **without** penalisation; it is common to fit spline APC in this manner. In `mgcv`, smoothing penalties are applied by default but are removed using the option `fx=TRUE`.

   (b) Penalised smoothing spline (PSS): This is the smoothing spline model fit **with** penalisation. The importance of penalisation will become clear in Section 4.

For all models, the *ad-hoc* choice of what linear slope to drop will be cohort (meaning the APC model is cross-sectional). Therefore, the models will contain age and period slopes and curvatures for all three temporal effects.

3.3 | Results

Identification issues due to the structural link are resolved by the *ad-hoc* forcing of one of the slopes to be zero. Due to this, comparisons between $h_\star$ and $\hat{h}_\star$ are inappropriate as the true effects do not have a zero linear trend. To compare the two sets of quantities, we construct identifiable functions of the true and estimated mean values.

Thus, we define modified true and estimated effects which take into consideration the intercept and structural link identifiability. In practise, first define the linear predictor for all APC combinations (including ones not present due to the structural link identification), then the adjusted true effects are calculated by subtracting the overall mean of the linear predictor from the marginal temporal effect of the linear predictor. For example, the true adjusted age effect for all three distributions is calculated,

$$h_A^+(a) = \begin{cases} \frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\mu_{apc} - \frac{1}{APC}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{c=1}^{C}\mu_{apc} & \text{Gaussian} \\ \frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\operatorname{logit}\left(\pi_{apc}\right) - \frac{1}{APC}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{c=1}^{C}\operatorname{logit}\left(\pi_{apc}\right) & \text{Binomial} \\ \frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\log\left(\lambda_{apc}\right) - \frac{1}{APC}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{c=1}^{C}\log\left(\lambda_{apc}\right) & \text{Poisson} \end{cases}.$$

The intercept identifiability is addressed in the true effects by subtracting the overall mean from the marginal of the linear predictor. As the structural link identifiability cannot be removed as with the intercept identifiability, it is consolidated into an 'average effect' of the remaining two terms. To see this explicitly and without the loss of generality, consider the Gaussian case where `offset` $= 0$ and `scale` $= 1$,

$$\begin{aligned} h_A^+(a) &= \underbrace{\frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\mu_{apc}}_{\text{marginal age effect}} - \underbrace{\frac{1}{APC}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{c=1}^{C}\mu_{apc}}_{\text{overall mean}} \\ &\propto \frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\left[h_A(a) + h_P(p) + h_C(c)\right] = h_A(a) + \underbrace{\frac{1}{PC}\sum_{p=1}^{P}\sum_{c=1}^{C}\left[h_P(p) + h_C(c)\right]}_{\text{Average period/cohort effect}} \end{aligned}$$

where the linear trends in period and cohort are consolidated into an average effect. For all distributions, the true age curvatures are found by de-trending the true effects

$$h_{A_C}^+(\boldsymbol{a}) = \left(\boldsymbol{I}_A - \boldsymbol{H}_A\right) h_A^+(\boldsymbol{a})$$

where $\boldsymbol{I}_A$ is an $A \times A$ identity matrix and $\boldsymbol{H}_A = [\mathbf{1} : \boldsymbol{a}]\left([\mathbf{1} : \boldsymbol{a}]^T [\mathbf{1} : \boldsymbol{a}]\right)^{-1} [\mathbf{1} : \boldsymbol{a}]^T$ is the hat matrix for an ordinary least squares fit of the age effect. To define the estimated effects, use the estimated linear predictor for all APC combinations instead of the true linear predictor, $\hat{\mu}_{apc}$, $\operatorname{logit}\left(\hat{\pi}_{apc}\right)$ and $\log\left(\hat{\lambda}_{apc}\right)$ for Gaussian, binomial, and Poisson, respectively, to define the marginal age effect and overall mean. The estimated curvatures are found using the estimated effects in the same manner as the true curvatures were. In all three distributions, the period and cohort effects and curvatures are analogous.

The results of the binomial simulation study are summarised in Figure 3. Each column refers to one of the temporal effects; age, period, and cohort from left to right. The first two rows show the estimated effect and curvature for each of age, period and cohort alongside their respective true effect and curvature. The latter two rows show the bias and mean square error (MSE) of the identifiable curvature terms. The bias and MSE for the effect at age a are $\frac{1}{K}\left[\sum_{k=1}^{K}\left(\hat{h}_{A_k}^+(a) - h_A^+(a)\right)\right]$ and $\frac{1}{K}\left[\sum_{k=1}^{K}\left(\hat{h}_{A_k}^+(a) - h_A^+(a)\right)^2\right]$, respectively and is analogous for period and cohort and the curvatures. The x-axis is labelled relative years since period is fixed to start at zero years and the cohort is relative to these.

The first row in Figure 3 shows the estimated and true full temporal effects. The shift and rotation in the estimates in comparison to the truth is because of the lack of identifiability in the full effects due to the structural link. The estimated full effects will change depending upon the arbitrary choices we make (such as how we define the orthogonal projection or the choice of linear slope to drop), but unless we have additional information, these estimates will (most likely) be different to the truth. To show how a more informed arbitrary choice of reparameterisation can give the impression of identifiability in the true effects, consider Figure S2 in Supplementary Material. Figure S2 shows the results of an APC model being fit to data generated without a cohort effect present. The additional information we have at our disposal (e.g., cohort is not present in the data generation) means we can make a more informed arbitrary choice (e.g., drop the cohort slope). Due to this, the estimated full effects are the same as the true effects. In reality, this external information would not have been available, and the true effects would not be identifiable.

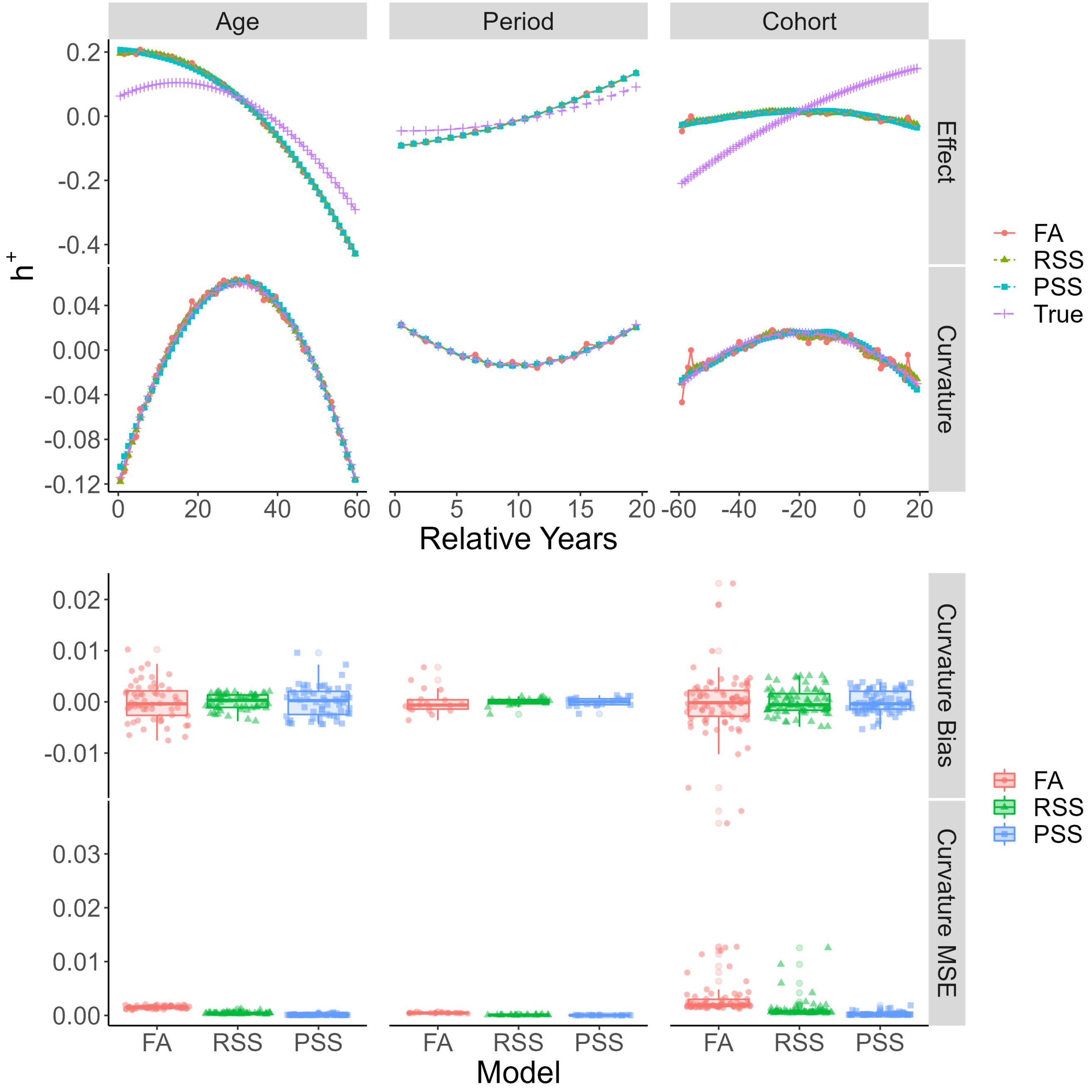


FIGURE 3 Simulation study results for equal interval, $M = 1$, binomial data generated when all three temporal effects are present. The FA, RSS and PSS models are the factor, regression smoothing spline and penalised smoothing spline models, respectfully. The first and second row are of the temporal effect and curvature plots for all models alongside the true values. The bottom two rows are the bias and MSE box plots for each model.

In contrast to the full effect, the temporal curvatures are identifiable, and this is shown in the second row of Figure 3 since the curvature estimates match the true curvatures. The FA curvature estimates are not as smooth as those from the RSS and PSS models. This is due to the fact the RSS and PSS models smooth between terms, with the PSS also penalising the "wiggliness" of the estimated functions. In addition, the added variability seen in the oldest and youngest cohorts (as these are the observations seen the least) is smoothed over in the RSS and PSS models and not in the FA model.

The bias and MSE are only displayed for the estimated curvatures as these are the identifiable component. Each model produces a set of curvatures that accurately estimate the true curvatures. The age bias for the RSS model is slightly larger than the other two but is still small (±0.05). The MSE further highlights the adequateness of each model. The behaviour of the PSS model for data generated in equal intervals is consistent, if not outperforming, what is expected from the well-known and well used FA and RSS models.

4 | UNEQUALLY AGGREGATED INTERVALS FOR AGE, PERIOD AND COHORT

Temporal data aggregated into intervals that match (e.g., five-year age, five-year period) are referred to as in ‘equal intervals’. If they do not match (e.g., five-year age, single-year period), the data is referred to as in ‘unequal intervals’. Providers of health and demographic data frequently release data that has been aggregated over multiple years. Even if collected in single years, it is common to be released aggregated over multiple years. This can be for several reasons, such as to preserve anonymity.

Unequally aggregated data can be considered in the simpler equal interval framework by collapsing over the lowest common multiple (LCM) of the intervals, LCM(age-years, period-years). Consider the following two cases $\text{LCM}(2, 1) = 2$ and $\text{LCM}(5, 3) = 15$. In the former, period is collapsed over two-groups leading to some information loss but potentially removes noise that obscures the true trend. In the latter, age is collapsed over three- and period over five-groups resulting in a larger amount of information lost. The more groups collapsed over, the fewer observations there are, inducing greater uncertainty in the parameter estimates.

4.1 | Curvature identification problem

Previously we have focused on the case where age and period are in equal intervals, $M = 1$. Table 2 shows how the cohort index varies when age is aggregated into an interval five-times larger than period, $M = 5$. Cohorts appear every fifth period, highlighted in blue, unlike in the equal interval case, Table 1, where cohorts appear every period.

TABLE 2 Cohort indexing for age-period data table where age is grouped $M = 5$ times larger than period. The cohort index is defined using $c = M \times (A - a) + p$ where $A = 8$ to fix the first cohort to be 1.

8	1	2	3	4	5	6	7	8	9	10
7	6	7	8	9	10	11	12	13	14	15
6	11	12	13	14	15	16	17	18	19	20
5	16	17	18	19	20	21	22	23	24	25
4	21	22	23	24	25	26	27	28	29	30
3	26	27	28	29	30	31	32	33	34	35
2	31	32	33	34	35	36	37	38	39	40
1	36	37	38	39	40	41	42	43	44	45
Age	1	2	3	4	5	6	7	8	9	10
	Period									

When fitting APC models to data in unequal intervals, the identification problem caused by the cyclic pattern has previously been represented in the transformed functions as an indicator function[5] and as a modulo function.[26] We use two M periodic functions $v_M(p)$ and $v_M(c)$ to represent a continuous time version of these. As v_M is periodic, $v_M(x + M) = v_M(x)$ and the subscript is used to denote the periodicity. The transformed functions that include all the identification issues are

$$\begin{aligned} \tilde{f}_A(a) &= f_A(a) - \text{const}_1 + \text{const}_3 M a, \\ \tilde{f}_P(p) &= f_P(p) + v_M(p) + \text{const}_1 + \text{const}_2 - \text{const}_3 p, \\ \tilde{f}_C(c) &= f_C(c) - v_M(c) - \text{const}_2 + \text{const}_3 c. \end{aligned} \tag{4}$$

In the equal interval case, the periodic functions were not considered explicitly. This is because, when $M = 1$ the periodic functions is constant for each period and cohort (for example, $\upsilon_1(p) = \upsilon_1(p+1) = \upsilon_1(p+2) = \dots$), so are implicitly considered in the const_2 term.

As previously discussed, the overall linear predictor is invariant to the inclusion of a constant (intercept) and linear term (structural link). Without loss of generality let $\text{const}_1 = \text{const}_2 = \text{const}_3 = 0$, the linear predictor for the transformed functions is

$$\begin{aligned} g\left(\tilde{\mu}_{ap}\right) &= \tilde{f}_A(a) + \tilde{f}_P(p) + \tilde{f}_C(c) \\ &= f_A(a) + \left[f_P(p) + \upsilon_M(p)\right] + \left[f_C(c) - \upsilon_M(p - Ma)\right] \\ &= f_A(a) + \left[f_P(p) + \upsilon_M(p)\right] + \left[f_C(c) - \upsilon_M(p)\right] \\ &= f_A(a) + f_P(p) + f_C(c) = g\left(\mu_{ap}\right) \end{aligned}$$

where $c = p - Ma$ and $\upsilon(m + M) = \upsilon(m)$ are used in the second and third lines, respectively. Thus, the linear predictor is invariant to the periodic function.

Now define the second derivatives of the two periodic functions as

$$\begin{aligned} \upsilon''_M(p) &\equiv \upsilon_{M_C}(p) \\ \upsilon''_M(c) &\equiv \upsilon_{M_C}(c). \end{aligned}$$

The previously identifiable second derivatives

$$\begin{aligned} \tilde{f}_A{}''(a) \equiv \tilde{f}_{A_C}(a) &= f_{A_C}(a) \\ \tilde{f}_{P_C}(p) &= f_{P_C}(p) + \upsilon_{M_C}(p) \\ \tilde{f}_{C_C}(c) &= f_{C_C}(c) - \upsilon_{M_C}(c) \end{aligned} \tag{5}$$

are no longer identifiable for period cohort due to the presence of υ_{M_C}. The period and cohort curvature identifiability issues mean these terms are no-longer estimable, as in the equal interval case; in the estimate, we cannot disentangle what is the "true" curvature from the periodic function. We will call this the "curvature identifiability problem".

4.2 | Resolving the curvature identifiability problem

Holford first acknowledge the curvature identification problem, calling it the micro-trend identifiability problem, when fitting factor models to APC data aggregated in unequal intervals. He described how the curvature identification problem produce an M-year cyclic pattern in the estimated temporal effects and proposed the use of smooth functions (smoothing splines) to model them.[9] This approach provides sufficient structure to smooth over the cyclic pattern caused by the curvature identification, but it does not address the problem itself. Additionally, both Heuer[11] and Carstensen[4] use smoothing splines when fitting an APC model to data aggregated in unequal intervals; however, both do not explore the curvature identifiability problem from the unequal intervals and whether continuous functions alone resolve this. An important practical aspect to consider when using splines is how to define the spline itself, i.e., what basis to use and how to define the knots for the basis. Because of this, each author includes their own recommendations on best practises (which vary slightly from another), but none have included a sensitivity analysis for these.

When fitting APC models to unequal data, the curvature identifiability problem means the period and cohort curvature functions are no longer estimable Eq.(5). Because of this, an infinite number of period and cohort estimates can be used to produce the same reparameterised linear predictor. Of those, if the function for period and cohort curvatures is approximating $\upsilon_M \neq 0$, the period and cohort estimates will display the arbitrarily large, cyclic pattern over M-years as described by Holford.[9] If $\upsilon_M \neq 0$ is approximated, the cyclic pattern in the period and cohort estimates is "wigglier" than when $\upsilon_M = 0$ is approximated. The PSS method we proposed in Section 2 has a penalty on the integrated square of the second derivative ("wiggliness") of the estimates. Therefore, an estimate for a function approximating $\upsilon_M \neq 0$ will have a larger integrated square of the second derivative and hence a larger penalty than one approximating $\upsilon_M = 0$. Consequently, the PSS method we propose will actively penalise the curvature identification issues; whereas, non-penalised methods will only smooth over them. A theoretical illustration of how the penalty function is alleviating the curvature identification problem can be found in the Supplementary Material.

4.3 | Simulation study

We now demonstrate this result empirically by repeating the simulation study from Section 3 with data in unequal intervals. We first generate the data in single-years and then aggregate, replicating the real-world practise of data being collected in single-year age and period format with aggregation occurring before the data is released. As is common in many epidemiology settings, we aggregate single-year age over five years (i.e., $\boldsymbol{M} = 5$).

The underlying single-year data are generated as described in Section 3. Once generated, the data is aggregated according to $\boldsymbol{p}$ and $\boldsymbol{a}'$, where $\boldsymbol{a}'$ is the M-year age vector of length $A' = A/M$. After aggregation, $\boldsymbol{p}$ and $\boldsymbol{a}'$ are used to define $\boldsymbol{c}'$, the cohort vector of length $C' = M \times \left(A' - 1\right) + P$ using $c' = p - a'$. In this simulation, we have A' such that it is an integer, i.e., each age group is aggregated over the same number of ages. If this is not the case, and A' is not an integer, the curvature identification problem would still be present.[9]

To get a true age effect that is comparable to the estimated age effect, we average the effect over every $\boldsymbol{M}$ distinct ages. Let a_i and a'_i be the i^{th} value in the vectors $\boldsymbol{a}$ and $\boldsymbol{a}'$, the true value of the age effect that is comparable to the aggregated estimated values is

$$h^+\left(a'_i\right) = \frac{1}{M} \sum_{m=-(M-1)}^{0} h^+\left(a_{[(i \times M)+m]}\right)$$

for $i = \{1, \dots, A' = \frac{A}{M}\}$. For example, we average the true age effect evaluated at 0.5, 1.5, 2.5, 3.5 and 4.5 to be comparable to the estimated effect at $a' = 2.5$.

Similarly, for cohort, average over every $\boldsymbol{M}$ cohorts (as age is aggregated in $\boldsymbol{M}$ years) and move along in single-year steps (as period is still single years). Therefore, let c_k and c'_k be the k^{th} value in the vectors $\boldsymbol{c}$ and $\boldsymbol{c'}$. The true value of the cohort effect that is comparable to the aggregated fitted values is

$$h^+\left(c'_k\right) = \frac{1}{M} \sum_{m=0}^{M-1} h^+\left(c_{k+m}\right)$$

where $k = \{1, \dots, C' = M \times \left(A' - 1\right) + P\}$. For example, average the cohort true effects at c =-59, -58, -57, -56 and -55 to be comparable to the estimated effect at $c' = -57$.

Once the age and cohort true values are aggregated, the bias and MSE will reflect the variability observed in the $M = 1$ simulations as well as the aggregation bias. As period is not changed, the expressions from Section 3 are used. The models fit are the factor (FA), regression smoothing spline (RSS) and penalised smoothing spline (PSS) defined in Section 3. The estimated effects $\hat{h}^+_\star$ and curvatures $\hat{h}^+_{\star_C}$ are calculated like Section 3 but with the vectors a' and c' for age and cohort; period is unchanged.

Figure 4 shows the results of the simulation study. Each column is one of the three temporal effects: the first two rows are the function plots of the estimated full effects and curvatures alongside the true functions of both and the bottom two rows are the bias and MSE for the curvatures.

The FA model displays the cyclic saw-tooth pattern which repeats every M-years (five-years) in both the full effects and curvatures for period and cohort, not age. The cyclic pattern in the period and cohort curvatures is due to the curvature identification problem. The period and cohort bias plots for FA model seem reasonable but this is due to the fact the cyclic pattern negating the overall bias. More telling is the difference between the FA models period and cohort MSE box plots and those from the RSS and PSS models; the FA box plot displays a general increase in the MSE values which themselves are more over-dispersed.

It is hard to differentiate between the results for the RSS and PSS models, with both seeming to provide adequate solutions to resolving the curvature identification problem. From the theoretical results, period and cohort curvature functions are not estimable, and the smoothest estimates come from a function that is approximating $v_M = 0$. Therefore, if the function chosen to approximate period and cohort is approximating $v_M = 0$, the estimates are the smoothest and there is no additional penalty being incurred from the added identification. In order to test this, we can perform a sensitivity analysis on the choice of function used to approximate the true functions in the first place.

4.4 | Sensitivity analysis

An important part of fitting APC models using splines is the basis and knot selection. Both Heuer[11] and Holford[9] use one type of spline basis and give specific recommendations of knot specification, with Holford recommending knots placing the knots every $\boldsymbol{M}$ distinct points and Heuer recommending every five distinct points. Carstensen extends both by calling for less specific placement of knots, instead recommending the knots scale with the number of distinct points.[4] Since all three have different

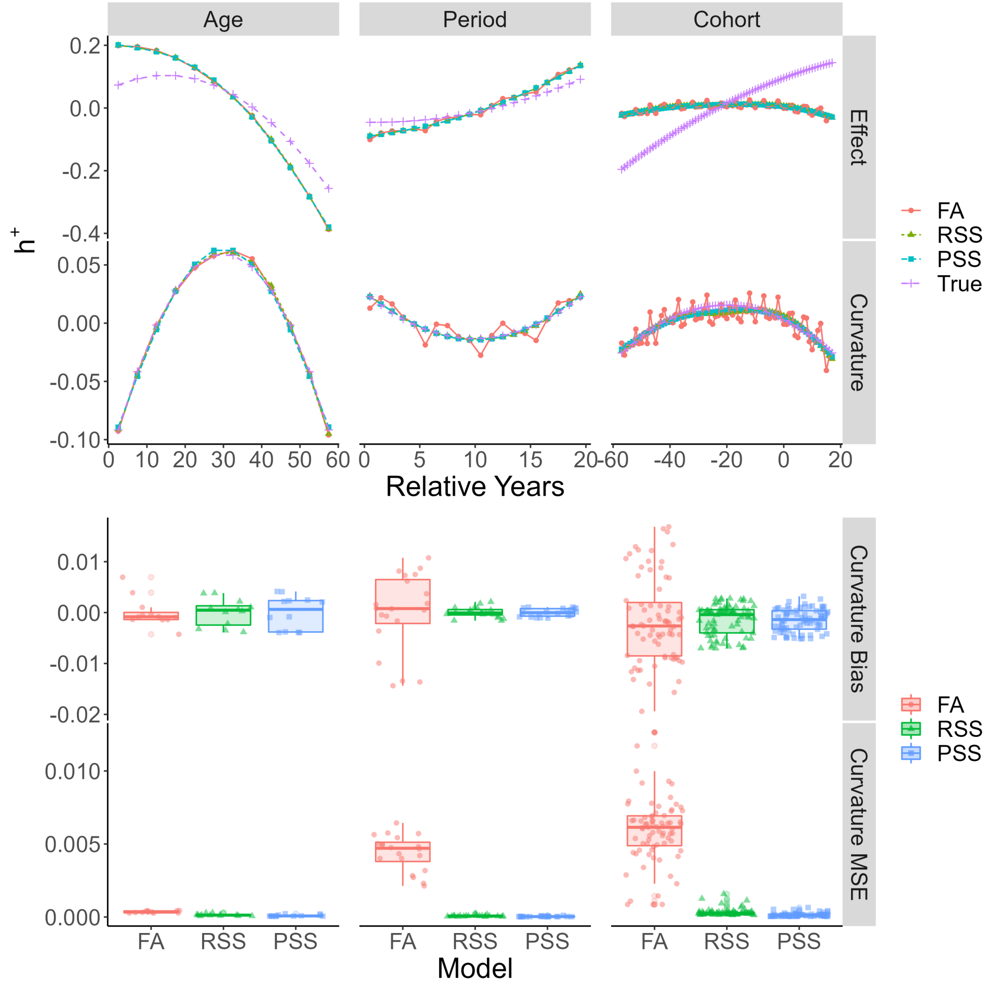
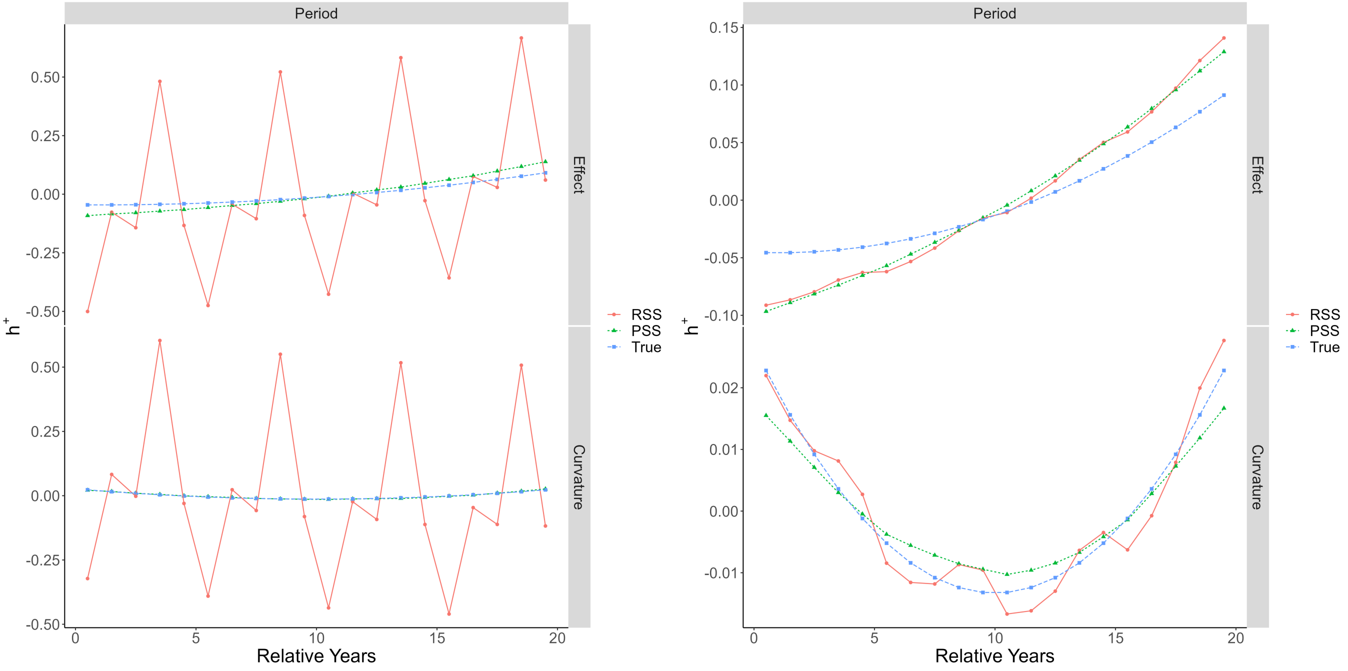


FIGURE 4 Simulation study results for unequal interval, $M = 5$, binomial data generated when all three temporal effects are present. The FA, RSS and PSS models are the factor, regression smoothing spline and penalised smoothing spline models, respectfully. The first and second row are of the temporal effect and curvature plots for all models alongside the true values. The bottom two rows are the bias and MSE box plots for each model.

recommendations of best practises when specifying the spline, we perform a sensitivity analysis to test the robustness of the RSS and PSS models results against the spline specification. Furthermore, this will show if the apparent alleviation of the curvature identification problem in both models is due to the methods working as intended or due to the choice in basis specification.

To test the robustness of the RSS and PSS model estimates to the spline specification, we perform a sensitivity analysis using two additional bases. The first additional basis is defined by more knots. Previously the number of knots was roughly 25% of distinct data points, and we increase this to be one less than the number of distinct data points. The second additional basis includes extra columns for a periodic function for period and cohort constructed using a cyclic cubic regression spline basis[16] alongside the normal cubic regression spline basis. The additional knots test the sensitivity to how the same basis is specified, and the additional periodic columns test the sensitivity to a different basis.

For conciseness, only the period effect results are shown in Figure 5, where Figure 5a and Figure 5b are the additional knots and additional periodic columns bases, respectively. Considering the curvature plots, both RSS estimates are different to one another and display a cyclic pattern. The three different patterns in the RSS estimates across the simulation study and sensitivity analysis show that the RSS results are sensitive to the basis specification. Furthermore, the inclusion of a cyclic pattern in each of the sensitivity analysis results means that the RSS model is not actually alleviating the curvature identification problem. In comparison, none of the estimates from the PSS model display the cyclic pattern; therefore, the PSS model is alleviating the curvature identification problem. For the PSS models additional periodic basis results, the penalisation does appear to over-smooth the function, but this can be attributed to the non-periodic basis elements also having the larger penalty applied to them.



(a) Additional knots.

(b) Inclusion of a periodic basis.

FIGURE 5 Simulation study results for the additional bases for unequal interval, $M = 5$, binomial data generated when all three temporal effects are present. Panel (a) shows the basis where the number of knots is increased to be one less than the number of distinct data points. Panel (b) shows the basis where there are additional periodic columns.

To summarise, the results from our simulation studies show the FA model is not suitable to model data unequally aggregated in any capacity. Additionally, the sensitivity analysis shows that without a penalty, results obtained from smoothing splines alone are not alleviating the curvature identification problem since the estimates are not robust to how the spline is specified. In contrast, the lack of the cyclic pattern in the PSS model estimates show a penalty function is successfully alleviating the curvature identification problem even when we specifically include cyclic elements in the basis. Therefore, we stress the importance and recommend the use of a penalty term on the estimates of the temporal curvature functions to provide robustness and give the user confidence that the curvature identification problem is addressed.

5 | APPLICATION

We now consider an application of the PSS model on UK all-cause mortality data downloaded from the Human Mortality Database (HMD).[27] This application is used to show that both the PSS model is appropriate for use with real-world data and to highlight how collapsing data that comes in unequal intervals into equal intervals can lead to incorrect, even contradictory, analysis.

The HMD contains the raw population and (all-cause) deaths of 41 countries around the world attained from a variety of national statistic offices. The data is not shareable but is free to download after registration. Data from the HMD was chosen as it is downloadable in single-year age and period which gives the freedom to aggregate it as required. The UK all-cause mortality data from the HMD comes in years 1922-2018 and age 0-110+. We take a subset of each and use years 1926-2015 and ages 0-99 to ensure equal groups when aggregating later. The HMD often receives data which is either already aggregated or contains missing values. Due to this, they fill in the missing information (using a method outlined in their method protocol[27]) which results in non-integer counts. For our models, we round the HMD values to the nearest integer.

Figure S9a in the Supplementary Material shows a heat map of the all-cause mortality in the UK for single-year age and period. For a fixed year and in the absence of cohort effects, the apparent age effects are the changes in mortality along the y-axis. For a fixed year and in the absence of cohort effects, the apparent age effects are the changes in mortality along the y-axis. For a fixed age-group and in the absence of cohort effects, the period effects are the change in mortality along the x-axis. The cohort effects reflect a combination of age and period effects and appear on the bottom left to top right diagonal. An example of a notable change for each effect is: for age, the mortality changing from extremely high to low in the first five-years of life when the year is fixed at 1926; for period, the drastic reduction in mortality for age groups 0-5 in more recent periods as oppose to earlier periods; and finally, a cohort effect is the yellow to red frontier in the top right diagonally increasing due to these cohorts having a better standard of living for the entirety of their life in comparison to cohorts before.

From the HMD data, three different data sets will be constructed: single-year age and period (1×1), five-year age and single-year periods (5×1) and five-year age and period (5×5). The 1×1 data represents the most informative data set where no aggregation occurs, and the 5×1 data reflects unequal interval data one might receive from a provider of health and demographic data. To show why collapsing unequal intervals into equal intervals is not a suitable method, we include the 5×5 data. This represents a case where one receives data in unequal form (i.e., in 5×1), but rather than address the curvature identification problem, the period group is collapsed over five years thereby producing a dataset with equal intervals. The first and last three rows of each data set can be seen in Table 3.

Aggregation	Age-Group	Year-Group	Population	Deaths
1×1	$[0, 1)$	$[1926, 1927)$	791373	59661
	$[0, 1)$	$[1927, 1928)$	763981	56260
	$[0, 1)$	$[1928, 1929)$	744778	53281
	···	···	···	···
	$[98, 99]$	$[2012, 2013)$	20340	7751
	$[98, 99]$	$[2013, 2014)$	20664	7711
	$[98, 99]$	$[2014, 2015]$	40198	15209
5×1	$[0, 5)$	$[1926, 1927)$	4026858	88081
	$[0, 5)$	$[1927, 1928)$	3888784	85596
	$[0, 5)$	$[1928, 1929)$	3773475	78393
	···	···	···	···
	$[95, 99]$	$[2012, 2013)$	90517	28900
	$[95, 99]$	$[2013, 2014)$	87777	27916
	$[95, 99]$	$[2014, 2015]$	187530	58471
5×5	$[0, 5)$	$[1926, 1931)$	18960706	411563
	$[0, 5)$	$[1931, 1936)$	17291084	329432
	$[0, 5)$	$[1936, 1941)$	16790532	277819
	···	···	···	···
	$[95, 99]$	$[2001, 2006)$	346142	114044
	$[95, 99]$	$[2006, 2011)$	416010	131290
	$[95, 99]$	$[2011, 2015]$	457192	142900

TABLE 3 UK all-cause mortality data aggregated in single-year age and period, five-year age and single-year period an five-year age and period.

To be consistent with the results displayed in the simulation study, we model the counts from a binomial distribution with a logit link function and drop the cohort slope during the reparameterisation. The model equation is

$$\text{logit}\left(\pi_{ap}\right) = \beta_0 + a\beta_{A_L} + p\beta_{P_L} + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c).$$

The number of knots used for each temporal effect is 10, 10 and 20 for age, period, and cohort, respectively. These are kept consistent across the models fit to all three data sets.

Figures S9b-S9d in the Supplementary Material show the predicted heat maps from each of the data sets. Since each of the predicted heat maps are in-line with the true heat map, the PSS model is appropriate to use for real-world applications. The difference in how the data is formatted can be seen by the pixel sizes in each of the figures and there are methods that can be used to generate smoother predicted heat maps[28]. Since the heat map is a graphical illustration of the linear predictor, which is invariant to the curvature identification problem, by considering them alone we are not able to tell if the curvature identification problem is being alleviated. To see this, we need to consider the temporal function themselves.

Figure 6 shows the smooth function of the curvatures estimated from each of the data sets. These estimates are not the same as the detrended temporal estimates $\hat{h}_{\star_C}$ from the simulation studies; they are the smooth functions of temporal curvatures themselves, $\hat{f}_{\star_C}$. The lack of a cyclic pattern in the 5×1 results confirm the curvature identification problem has been addressed by the penalisation.

The smooth functions of curvatures represent the rate of change in a given direction. For example, the steep positively increasing half of the cohort curvature estimate reflects large improvements (large changes) in mortality in comparison to prior cohorts rather than an increase in mortality. Furthermore, the steep negatively decreasing half does not reflect a reduction in mortality rates but rather the improvements in mortality from cohort to cohort being smaller than before. In Figure S9a in the Supplementary Material, these changes can be seen. The prominent diagonal frontier between the light blue and dark blue for ages 10-30 and years 1930-1960 is steeper than the frontier for 1960-present in the same age range. This means for the same ages, the apparent cohort effect reducing mortality is less pronounced. This could be from advances in living standards slowing down for the latter half of the 1900s onwards.

Given age is aggregated over five in the 5×1 and 5×5, the two sets of estimates of smooth functions are imperceptibly different to one another, hence the appearance of only two curves in the age column of Figure 6. Both aggregated age estimates follow roughly the same trend as the un-aggregated estimates. The effect of the aggregation is clear: the more drastic changes in mortality are not captured in as much detail when aggregating. Given the slower rate of change in the cohort estimates, it is no surprise the three sets of cohort estimates are extremely similar. Each of the three functions follow a similar path, reach similar peaks, and have similar start and end points.

The difference between the 5×1 and 5×5 is apparent in the estimates for the period smooth functions. At times of large change, the estimates from the 5×5 do not capture the full extent of change (e.g., the 1930s peak and 1950s trough) and have conflicting estimates (fluctuations in the 2000s). In comparison, the 5×1 model, which does not rely on collapsing to resolve added identification, follows the more informative 1×1 estimates extremely well.

Clearly, aggregating over groups loses information. The difference between the 1×1 and the other two estimates for age smooth functions show this. To then collapse the aggregated data further, from 5×1 to 5×5, loses even more information and reduces the explanatory power of the model. When data does not come in equal intervals, there is still substantial information present to give detailed representation of how mortality changes over time. This application clearly demonstrates that further collapsing to avoid complications negatively impacts the explanatory power of the model, which will impact the reason the model is being fit in the first place (evaluating interventions, policy change, analysis, etc.).

Being able to capture the larger changes in mortality for a given effect is one of the most important aspects of mortality modelling. Gaining insight into what causes these changes is helpful to understanding whether similar changes will happen again and if so, will interventions help in any way. The APC penalised smoothing spline model on unequal data produces estimates in-line with the richer data, highlighting the importance for a method that can handle data in any format.

6 | CONCLUSION

In this paper, we conducted a simulation study and sensitivity analysis to investigate the use of penalised smoothing splines (PSS) on the well-known curvature identification problem that arises when fitting APC models to data tabulated in unequal intervals. The proposed method was compared to two different implementations of the same reparameterisation scheme, a factor (FA) version[2] and a regression smoothing spline (RSS) version of the model. The result of the simulation study and sensitivity analysis for data in unequal intervals showed the PSS model is robust at alleviating the curvature identification problem unlike the currently used FA and RSS methods. Further benefits of an APC model that can appropriately handle unequal data were described during an application to UK all-cause mortality data from the HMD.

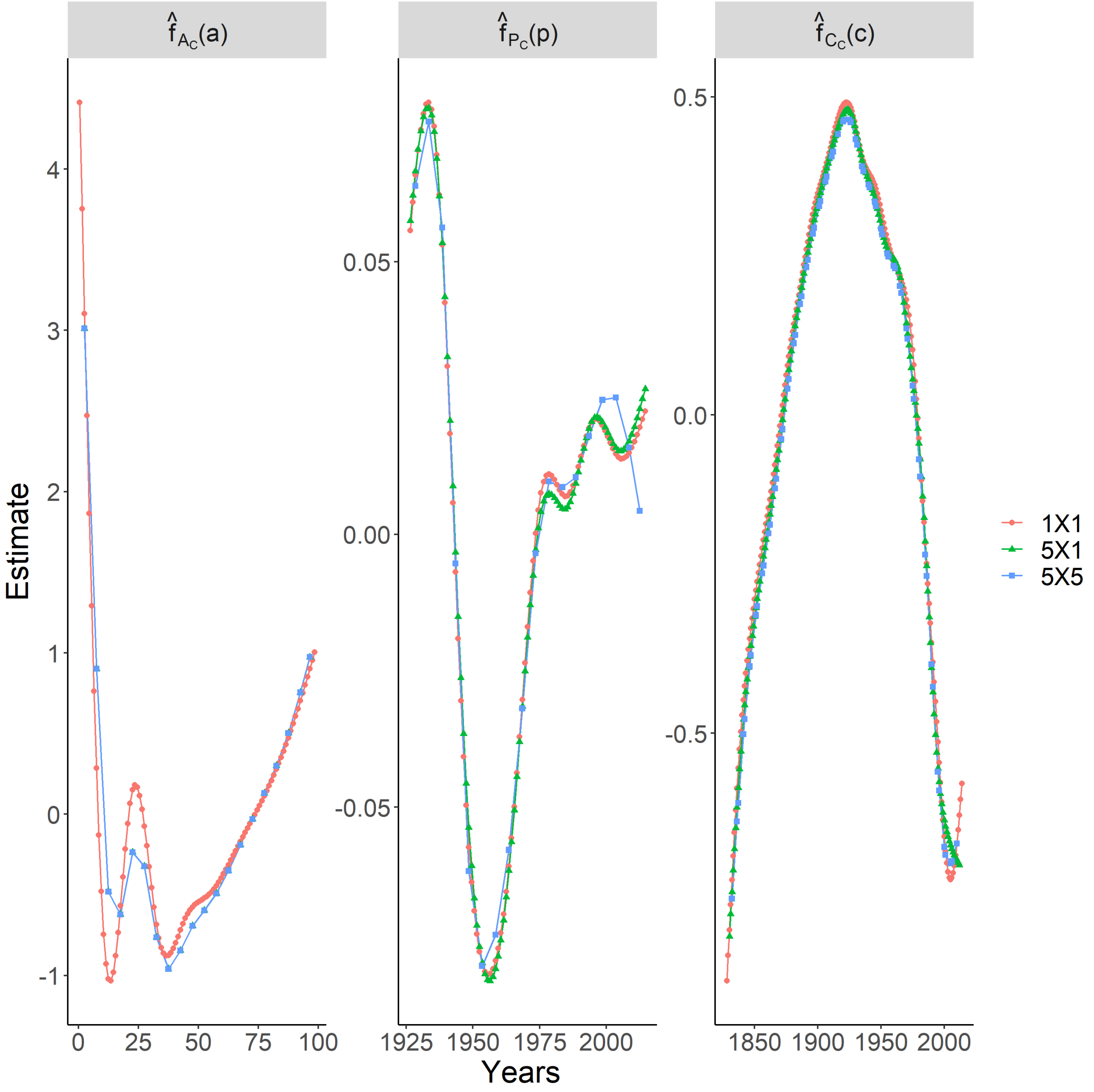


FIGURE 6 UK all-cause mortality mortality fitted smooth curvatures for models fit to data aggregated in single-year age and period, five-year age and single-year period and five-year age and period.

The simulation study for data aggregated in equal intervals demonstrates the PSS model resolves the usual APC structural link identification problem. The unequal intervals simulation study compared the suitability of our model and those from the literature at addressing the curvature identification problem. The cyclic pattern in the FA model estimates highlighted the issue and showed this model is not appropriate in any capacity. The results from a sensitivity analysis show the RSS model does not provide a robust solution to the problem; whereas, the PSS model does. Consequently, we strongly recommend the use of a penalty to give robustness to, and confidence in, the results when fitting APC models to data that comes in unequal intervals.

We demonstrated with penalised smoothing splines it is essential to include a penalty when fitting APC models to data that is aggregated in unequal intervals. An alternative method to including a penalty on the estimates is to use a smoothing prior in a Bayesian paradigm, such as a random walk prior[26] or a Gaussian process prior.[29] These methods have parallels between our method due to the correspondence between penalised smoothing splines and stochastic processes[30,31] and we believe these provide suitable solutions. However, more work (which we have in progress) still needs to be done to fully explore the appropriateness of smoothing priors within the context of the curvature identification problem for APC models.

We consider APC methods that only reparameterise into a curvature (or equivalent) component, but there are alternative reparameterisations. For example, the curvature can be further split into an interpretable quadratic term, which gives additional insight into how fast the temporal trends are changing, and higher-order terms.[32] However, this is yet to be considered with respect to data that comes in unequal intervals..

An extension of this framework is to include forecasts. Forecasting with an APC model in a health setting is useful when updating policies and allocating resources. Consider the unidentifiable APC model where we are predicting h periods into the future

$$g\left(\mu_{a,p+h}\right) = f_A(a) + f_P(p+h) + f_C(c+h)$$

where $c = M \times (A - a) + p$. Forecasts depend on estimates, from the data, of the period and cohort functions to be projected h steps ahead. When the individual temporal trends are not of interest, forecasts can be made from the above unidentifiable model. However, forecasting is more likely to be used to answer questions such as response of a given age-group over the coming years; this requires knowledge of the temporal trends. Therefore, the best practise is to perform forecasting based on invariant forecasting functions,[3] which in our proposal are the temporal curvatures.

In this paper, we consider all the temporal intervals in the unequal interval data to have a constant width, and do not consider when the data comes in the format of non-constant unequal intervals. An example of non-constant unequal intervals is the weekly period with five-year age groups from the ONS;[7] the first age group is split into $[0, 1)$ and $[1, 4)$. An additional example comes from the DHS,[8] where when modelling under-five mortality, it is common to aggregate the monthly ages into $[0, 1)$, $[1, 12)$, $[12, 24)$, $[24, 36)$, $[36, 48)$ and $[48, 60)$. Both the ONS and the DHS split age like this to better capture the changes in mortality as the first year and month are extremely different to the rest. Other APC models have been extended to incorporate covariates in space[33,34] and such extensions of our work would be possible, too. The biggest challenge that will be encountered when extending our proposed APC reparameterisation is keeping track of identifiable terms, especially when considering, for example, within covariate temporal trends. Computational challenges can arise using splines to approximate functions for large datasets (such as the DHS data) or for spatial extensions; therefore, a more efficient smooth approximation may need to be considered.

ACKNOWLEDGMENTS

The authors would like to thank Professors Christopher Jennison and Jon Wakefield for their helpful comments on earlier drafts. Furthermore, the authors are grateful for the insightful suggestions of the four referees and associate editor who helped improve this manuscript.

Conflict of interest

The authors declare no potential conflicts of interest.

References

1. NHS . *Health Survey for England*. National Health Service; https://www.digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england: 2020. Accessed August 2021.

2. Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983; 39(2): 311–324.

3. Kuang D, Nielsen B, Nielsen JP. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* 2008; 95(4): 979–986.

4. Carstensen B. Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 2007; 26(15): 3018–3045.

5. Smith TR, Wakefield J. A review and comparison of age–period–cohort models for cancer incidence. *Statistical Science* 2016; 31(4): 591–610.

6. ONS . *Deaths registered by single year of age, UK*. Office For National Statistics; https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathregistrationssummarytablesenglandandwalesdeathsbysingleyearofagetables: 2020. Accessed August 2021.

7. ONS . *Deaths registered weekly in England and Wales, provisional*. Office For National Statistics; https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales: 2020. Accessed August 2021.

8. USAID . *Demographic and Health Surveys*. United States Agency for International Development; http://www.dhsprogram.com: 2019.

9. Holford TR. Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine* 2006; 25(6): 977–993.

10. Held L, Riebler A. A conditional approach for inference in multivariate age-period-cohort models. *Statistical Methods in Medical Research* 2012; 21(4): 311–329.

11. Heuer C. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics* 1997: 161–177.

12. Mason KO, Mason WM, Winsborough HH, Poole WK. Some methodological issues in cohort analysis of archival data. *American Sociological Review* 1973: 242–258.

13. Fienberg SE, Mason WM. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology* 1979; 10: 1–67.

14. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.

15. Rutherford MJ, Lambert PC, Thompson JR. Age–period–cohort modeling. *The Stata Journal* 2010; 10(4): 606–627.

16. Wood SN. *Generalized additive models: An introduction with R*. Chapman and Hall/CRC. Second ed. 2017.

17. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Routledge . 1990.

18. Flegal KM, Carroll MD, Ogden CL, Johnson CL. Prevalence and trends in obesity among US adults, 1999-2000. *The Journal of the American Medical Association* 2002; 288(14): 1723–1727.

19. Mokdad AH, Ford ES, Bowman BA, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *The Journal of the American Medical Association* 2003; 289(1): 76–79.

20. Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 1999-2004. *The Journal of the American Medical Association* 2006; 295(13): 1549–1555.

21. Cole TJ, Power C, Moore GE. Intergenerational obesity involves both the father and the mother. *The American Journal of Clinical Nutrition* 2008; 87(5): 1535–1536.

22. Kuh D, Shlomo YB. *A life course approach to chronic disease epidemiology*. Oxford University Press. Second ed. 2004.

23. Luo L, Hodges JS. Block constraints in age–period–cohort models with unequal-width intervals. *Sociological Methods & Research* 2016; 45(4): 700–726.

24. Fannon Z, Monden C, Nielsen B. Modelling non-linear age-period-cohort effects and covariates, with an application to English obesity 2001–2014. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2021; 184(3): 842–867.

25. Rosenberg PS, Check DP, Anderson WF. A web tool for age–period–cohort analysis of cancer incidence and mortality rates. *Cancer Epidemiology and Prevention Biomarkers* 2014; 23(11): 2296–2302. Available at https://analysistools.cancer.gov/apc/.

26. Riebler A, Held L. The analysis of heterogeneous time trends in multivariate age–period–cohort models. *Biostatistics* 2010; 11(1): 57–69.

27. HMD . *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany); www.mortality.org: 2020. Accessed May 2022.

28. Chien LC, Wu YJ, Hsiung CA, Wang LH, Chang IS. Smoothed Lexis diagrams with applications to lung and breast cancer trends in Taiwan. *Journal of the American Statistical Association* 2015; 110(511): 1000–1012.

29. Chernyavskiy P, Little MP, Rosenberg PS. Correlated Poisson models for age-period-cohort analysis. *Statistics in medicine* 2018; 37(3): 405–424.

30. Wahba G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)* 1978; 40(3): 364–372.

31. Speckman PL, Sun D. Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* 2003; 90(2): 289–302.

32. Rosenberg PS. A new age-period-cohort model for cancer surveillance research. *Statistical methods in medical research* 2019; 28(10-11): 3363–3391.

33. Etxeberria J, Goicoa T, López-Abente G, Riebler A, Ugarte MD. Spatial gender-age-period-cohort analysis of pancreatic cancer mortality in Spain (1990–2013). *PloS one* 2017; 12(2): e0169751.

34. Chernyavskiy P, Little MP, Rosenberg PS. Spatially varying age–period–cohort analysis with application to US mortality, 2002–2016. *Biostatistics* 2020; 21(4): 845–859.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article. Furthermore, code to replicate the simulation studies and application analysis can be found at https://github.com/connorgascoigne/Unequal-Interval-APC-Models.

□