Modelling of spatio-temporal disease rates using age-period-cohort methods

submitted by

Connor William Sydney Gascoigne

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

September 2022

COPYRIGHT NOTICE

DECLARATION OF ANY PREVIOUS SUBMISSION OF THE WORK

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Signature of Author.....

Connor William Sydney Gascoigne

DECLARATION OF AUTHORSHIP

I am the author of this thesis, and the work described therein was carried out by myself personally in collaboration with my supervisor and those mentioned below.

Chapter 2 is reproduced from a manuscript undergoing revisions from Statistics in Medicine peer review.

Chapter 3 is reproduced from a prepared manuscript.

Chapter 4 is reproduced from a manuscript in collaboration with Dr Theresa Smith from the University of Bath and Dr Johnny Paige from Norwegian University of Science and Technology and Professor Jon Wakefield from the University of Washington. The work was conducted and written by myself with my collaborators providing feedback and expertise on their relevant field.

Signature of Author.....

Connor William Sydney Gascoigne

ABSTRACT

Age-period-cohort (APC) models are used to analyse a variety of different health and demographic related outcomes. When including all three temporal trends into one model, there arises the well-known identification problem due to the structural link between the temporal trends (given two, the third can be calculated). Previous methods to resolve this problem focus on defining a model based off identifiable quantities. Most of the literature focusses on the case when APC models are fit to data aggregated in equal intervals (age and period widths). This may be, in part, due to the added complication that arises when fitting APC models to data in unequal intervals which causes a cyclic pattern in any estimates of the temporal trends.

We first show that when an APC model is fit to data in unequal intervals, the previously identifiable terms, used to define a model that resolves the problem created by the structural link, are no longer identifiable. Using penalised smoothing splines, we show how the novel inclusion of a penalty on the previously identifiable terms resolves any problems that arise when fitting an APC model to data aggregated in unequal intervals and conclude the necessity of a penalty to provide a suitable solution. We demonstrate the suitability of our proposed model using theoretical and empirical results.

Subsequently, we highlight the key information that links our proposed method to a class of APC models that utilise smoothing priors in a Bayesian paradigm. We give a full consideration of the problems that arise when fitting APC models to unequal interval data and how smoothing priors can be used to resolve them. Using theoretical and empirical results, we show that the smoothing prior models are performing a similar penalisation to that of a penalised smoothing spline, which we had previously shown to be a suitable solution, concluding their suitability is due to this penalty.

We conclude with a novel application of an APC model to under-five mortality rates (U5MR)

in Kenya 2006 to 2014. Previous methods to model U5MR include temporal terms for age and period, but not cohort. We extend the APC model to be suitable for the application by including spatial, spatio-temporal and strata specific terms alongside the terms for age, period, and cohort. The results indicate the inclusion of cohort is important to producing smooth fitting subnational estimates, an important goal for U5MR modelling which suffers from unstable estimates due to sparse data.

DEDICATION

In memory of

Enid Couling

1930 - 2021

ACKNOWLEDGEMENTS

There are many people I owe thanks to for their continued support and encouragement over the course of my PhD. As expected (and extremely cliché), without these individuals, the following pages would have been much harder (or not even possible at all!) to write.

First and foremost, I owe an endless amount of thanks and gratitude to my supervisor, Theresa Smith. From first introducing me to spatial statistics and public health, to constantly being my inspiration throughout, and finally, to helping start my career as a researcher. Your endless support and patience have been phenomenal. Thank you for your guidance, friendship, and always pushing me to be the best version of myself. I will truly miss our meetings where we discuss vital issues such as, where to go for the best beer in town, if this piece of work really is 'no drama', and, most importantly, whether the term is 'pernickety' or 'persnickety'.

To my examiners, Andrea Riebler and Julian Faraway, thank you for kindly accepting to review my thesis and making the viva an amazing experience that I will not forget. I would also like to thank Jon Wakefield for hosting me on a research visit to the University of Washington and welcoming me into his working group, STAB. To the STABees, thank you for the fruitful discussions and being expert tour guides on my visit.

An important part of my time at Bath has been the institution known as level five. Whether you were a member, or a passer-by, each of you I had the pleasure to spend time with were brilliant and provided an amazing working environment that I will truly miss. From Ben and Seb in the early days to Alex and Liam in the latter days, thank you all for the endless entertainment. Special thanks go to Emiko, for always being there with coffee, porridge, and wise words no matter the problem; Jen, for the weekend adventures and terrible puns; and Matt, for broadening my horizons and always entertaining my terrible ideas.

To those outside of Bath that have influenced and supported me before and during the PhD, I

am forever grateful. Alex, Dan, and Jake, thank you for always being ready with a beer, terrible chat, and endless dubs; and Josh, Matt, Sam, and Samir, thank you for your continued friendship throughout. Finally, my deepest gratitude goes to my family, all of whom have supported me through the good and bad times even when they had no clue what I was on about (which, I suspect, was most of the time!). My parents, Bill and Sarah, my sister, Shelagne, as well as Carl, Melanie, Mike, and Susan, this journey would not have been possible to start, let alone finish, without you!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The outcome (incidence and/or death) for a given health concern changes in time for various reasons. Understanding the influence these factors have on the outcome of population health is of great concern. However, measuring such factors directly can be extremely difficult, not least due to limits to data access. For example, if mortality increases following the introduction of a new disease, was the increase in deaths due to the new disease alone? Or was it due to lack of availability to appropriate care? Or a combination of the two? To avoid these issues, rather than look at particular factors, we can look at the time scale along which they both operate. As the health concerns change in time because of these factors, it is natural to consider temporal scales to describe and predict changes in health outcomes. The aim of this thesis is to use the so-called age-period-cohort (APC) models to model the relationship between certain temporal scales and health outcomes. In particular, we consider the technical issues that arise for APC models when the data used comes in non-uniform format.

There are several different temporal scales that can be considered when looking to describe changes in health outcome, but the three most influential temporal scales are age, period, and cohort. An age effect is a measure of attrition and physical changes on the body as we get older. Age is often considered the most important temporal effect due to the influence it has in most disease incidence and mortality rates. Period effects are short term exposures (e.g., new disease and new treatments) that have an immediate, lasting impact. Cohort effects (commonly birth cohort) are long term exposures (e.g., childhood pollutant exposure and smoking views) that a group of people who are of similar age encounter as they move through life together.

The datasets needed to analyse temporal scales alone are normally much simpler than those that include other factors relating to health outcomes. For example, consider the question we raised earlier: was an increase in deaths due to the introduction of a new disease, the lack of care, or a

combination of the two. To study this, we need a way to measure the new disease (e.g., using a binary variable to indicate if the individual has the disease or not) and a way to measure lack of care (e.g., a scale based on a scoring metric). This information is hard to come by, increases the complexity of the dataset, and may require additional special licenses to access. Alternatively, since both are short term exposures and have an immediate effect on the health outcome, they are captured by period, which is readily available, simple, and often does not need additional licensing to access.

When reporting health outcomes along temporal scales, data is often tabulated in a two-dimensional array where each axis represents a different time scale. Commonly, age and period are the chosen two time scales as we always know the age of the individual when the outcome occurs and the period (year) of the outcome. Figure 1-1 shows an example of how tabulated data is released from most providers of health and demographic data with age and period on the vertical and horizontal axes, respectively. The blue rectangle represents a five-year old and the yellow rectangle represents the period 2005. When age and period are considered together, the (birth) cohort can always be calculated since there is a linear relationship between the three, $cohort = period - age$. In Figure 1-1, the birth cohort of the five-year-old in 2005 is the year 2000, and this is the green trapezoid. Whilst we have chosen to keep the age and period scales to be one-year increments and equal, the increments can be any size and non-equal. Even with these alternative tabulations, given any two of age, period, and cohort, the third can always be calculated.

It is common for most health and demographic registry data to come in the tabulation shown in Figure 1-1. Therefore, it is common to use APC models to explore the dynamics between age, period and cohort and the health outcome since they capture all three temporal effects (trends) together. A search on PubMed for "age-period-cohort" showed that in the last year alone, APC models have been used to explore the relationship between the three temporal trends and chewing ability (Kim and Kawachi, 2021), stroke mortality (Cao et al., 2021), suicide trends (Chen et al., 2021; Martínez-Alés et al., 2021), alcohol consumption (Baburin et al., 2021), firearm homicide and suicide (Haviland et al., 2021), mortality of Type II diabetic kidney disease (Wu et al., 2021), aerodigestive tract and stomach cancer mortality (Kuzmickiene and Everatt, 2021), dementia (Kolpashnikova, 2021) and cervical, ovarian and uterine cancer mortalities (Wang et al., 2021) amongst many other health concerns. The search highlights the range of health-related outcomes that APC models can be used for.

As mentioned previously, there is a linear dependence between the three temporal trends, which we will refer to as the 'structural link'. The structural link is inherent in all data tabulated with two of the three temporal scales on the x and y axis, respectively, and when a model containing all three temporal trends is fit to data with this link, it causes what is known as the 'structural link identification problem' (Mason et al., 1973; Fienberg and Mason, 1979). To highlight the structural link identification problem, we must first consider an APC model. Let $a = 1, \ldots, A$ and $p = 1, \ldots, P$ be the number of distinct age and period groups. The true cohort is calculated

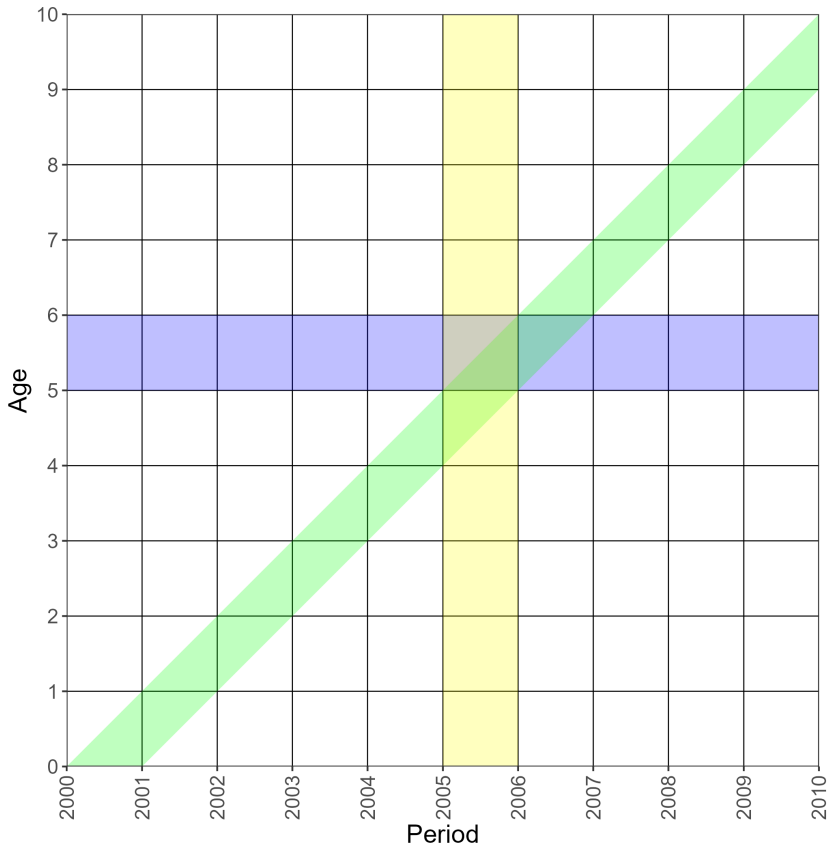


Figure 1-1: An example tabulation for a given health outcome measure by temporal scales.

using $c = p - a$, but if we want the earliest cohort to be indexed from one, we can use the equivalent $c = M \times (A - a) + p$ where M is the ratio of age interval to period interval. When $M = 1$, we refer to this as equal interval data since the number of ages and periods aggregated over is the same (e.g., single-year age and period or five-year age and period) and when $M \neq 1$, we refer to this as unequal interval data (e.g., $M = 5$ could indicate five-year age and single-year period).

A (continuous) APC model that captures all three temporal trends is,

$$g(\mu_{ap}) = \eta_{ap} = f_A(a) + f_P(p) + f_C(c) \tag{1.1}$$

where $g(\cdot)$ is the link function, $\mu_{ap} \equiv \mathbb{E}[y_{ap}]$ is equivalent to the expected value of the response, η_{ap} is the linear predictor and f_A, f_P and f_C are the smooth functions of age a, period p and cohort c.

First consider the simple case of equal interval data. The structural link identification problem means the linear predictor is invariant to the inclusion of a linear term alongside each of the

temporal terms. For example, consider the following set of transformed temporal functions,

$$\tilde{f_A}(a) = f_A(a) + \text{const}_3 a$$
$$\tilde{f_P}(p) = f_P(p) - \text{const}_3 p$$
$$\tilde{f_C}(c) = f_C(c) + \text{const}_3 c.$$

By using $c = p - M \times a$ with $M = 1$, it is easy to confirm the linear predictor is invariant to the addition of the linear trend i.e., $\tilde{\eta}_{ap} = \eta_{ap}$. There are many recognised solutions to the structural link identification problem for data aggregated in equal intervals which we will discuss in due course.

When data comes aggregated in unequal intervals, the structural link identification problem is still present, but there are additional issues that cause a cyclic (saw-tooth) pattern in the temporal estimates for period and cohort with a periodicity of M (Holford, 2006; Held and Riebler, 2012). Since the additional issues are displayed by the way of a cyclic pattern repeating every M period and cohort, we choose to represent them by a M-periodic function which we can include in the transformed period and cohort functions. Therefore, let v_M be an M periodic functions such that $v_M(x + M) = v_M(x)$, and the subscript denotes the periodicity. A similar characterisation of the problem has been made for discrete time where instead of a continuous periodic function, an indicator function is used to represent the addition and subtraction of a constant to every M^{th} period and cohort, respectively (Smith and Wakefield, 2016). The transformed functions can now be expressed,

$$\tilde{f_A}(a) = f_A(a) + \text{const}_3 M a$$
$$\tilde{f_P}(p) = f_P(p) - \text{const}_3 p + v_M(p)$$
$$\tilde{f_C}(c) = f_C(c) + \text{const}_3 c - v_M(c).$$

By using $c = p - M \times a$ and $v_M(x + M) = v_M(x)$, the linear predictor can be shown to be invariant to the both the linear trend and the cyclic function. We refer to these additional problems that arise when fitting an APC model to data unequally aggregated as the 'curvature identification problem'. The reason for this will be made clear later in this chapter.

Since the curvature identification problem is an additional problem on-top of the pre-existing structural link identification problem, most of the APC literature only considers models fit to equal interval data for simplicity. This contrasts with the fact that most providers of health and demographic data often release data in unequal intervals. For example: the UK's office for national statistics (ONS) regularly release data in unequal intervals with five-year age intervals and weekly period intervals being an example (ONS, 2020b); the Demographic and Health Surveys (DHS), a provider of survey data for low-and-middle income countries, release data with monthly ages and yearly period (USAID, 2019); and cancer data from the United Kingdoms

(UK) National Health Service (NHS) comes in single-year period and five year ages[1]. There are multiple reasons data are released aggregated in unequal temporal intervals, for example: to maintain confidentiality where low counts occur when data is heavily stratified or when there are a lot of zero counts; the dataset could be large with many covariates and to help users analyse the data, the providers may release it aggregated; or simply, that is how the data was collected.

We will contribute to the field of APC modelling by providing a thorough description of the problems involved when fitting APC models to data in unequal intervals allowing us to critique current popular methods. Additionally, based upon our critique, we will design a new APC model that is appropriate and apply this to a novel application where the APC model has not been used due to the data coming tabulated in unequal intervals.

1.1 Aims and contributions of this thesis

The aim of this thesis is threefold: (i) identify what makes an APC model resolve the curvature identification problem and critique the appropriateness of current methods as to whether they do this; (ii) based on the critique of the current methods, define an APC model that is appropriate for data in unequal intervals; and (iii) we provide a novel application of the model we propose to an important issue in public health.

Development of a new APC model and understanding of what makes an APC model appropriate for unequal interval data is the focus of Chapter 2. We extend the current methods by novelly including a penalty on the estimates of the reparameterised continuous functions of age, period, and cohort. We test the suitability of our model using both simulated and real-world data that is aggregated in both equal and unequal intervals. In addition, we highlight the strengths of our model by comparing the results found from our proposed models against models used in the literature. Having identified a suitable mechanism for alleviating the curvature identification problem, the focus of Chapter 3 is to extend this to include a wider range of models that can be used when fitting APC models to data in unequal intervals. To show the importance of the work we have done in this thesis, in Chapter 4 we use our model in a novel application to an important problem in public health that has not been able to include all three temporal effects together because of the structural link and curvature identification problems.

This thesis is aimed at both theoretical and practical users of APC models. Because of this, we include both theoretical and empirical justification to the model we develop and the claims we make. For the theoretical work, we aim to not only describe what we are doing, but why we are doing this in a way that is understandable for all. For the empirical work, we use both simulated and real data and, in each case, perform appropriate analysis to justify the

[1] https://www.cancerdata.nhs.uk/

model. When we define our model in Chapter 2, we give a rigorous explanation of both the theoretical and practical methods that occur to fully justify our novel extension, and why this model appropriately addresses the curvature identification problem. In Chapter 3, we identify the key theoretical properties between the models used in Chapter 2 and those used in Chapter 3 to expand the type of APC models we believe are appropriate for unequal interval data. To ensure this work is accessible for APC modellers of all abilities, we adopt a more heuristic and pragmatic approach to the justification and explanation in Chapter 3. For each method of explanation, we use the empirical results to provide a data-driven explanation and keep the practical motivation of the model in mind.

The PubMed search highlighted the extensive range of public health applications that APC models are used in. To ensure the model we develop is suitable for application in public health, the data used in all empirical results are either explicitly from, or derived from, a public health setting. The simulated datasets in Chapters 2 and 3 are based off obesity survey data from the UKs NHS (NHS, 2020). In Chapter 2, we use the UKs all-cause mortality from the Human Mortality Database (HMD) (HMD, 2020) as a case-study. In Chapter 4, we produce a full analysis using our proposed model to produce subnational estimates and predictions of under-five mortality rates (U5MR) using survey data from the DHS (USAID, 2019).

1.2 Identification of an APC model

An identification, or more appropriately lack of identification, problem in a statistical model is where the parameters of the model cannot be defined from the data alone. An easy example of this is an equation such as $\alpha x + \beta y = 10$, where we cannot define the parameters (α and β) from the data (the right-hand side of the equality) alone. Because of this, an infinite set of parameters can be determined from the model fitting process meaning it is impossible to perform any substantial analysis without steps being taken to ensure the identifiability of the model. Consequently, we say the parameter space is overparameterised. A model is identifiable if all the parameters in the model are estimable.

In general, the APC model defined in Eq.(1.1) is unidentifiable and for any type of data (equal or unequal intervals) has two identification problems. The first identification problem is common for all models that have more than one term (e.g., the smooth functions f) included (McCullagh and Nelder, 1989) and we call this the 'overall level identification problem'. The second identification problem is specific to the APC model and is due to the structural link $c = p - M \times a$ between the three temporal covariates (Mason et al., 1973; Fienberg and Mason, 1979) which we call the 'structural link identification problem'.

Since the overall level identification problem is present in any statistical model with more than

one term, it has been considered thoroughly and there are several prescribed solutions (McCullagh and Nelder, 1989). Similarly, the structural link identification problem in APC models has been considered by numerous authors in several different settings (Mason et al., 1973; Fienberg and Mason, 1979; Holford, 1983; Clayton and Schifflers, 1987; Osmond and Gardner, 1982; Carstensen, 2007; Kuang et al., 2008), and a general trend in acceptable solutions is clearly defined. We will cover the accepted solutions to both the overall level and structural link identification problems in the latter part of this Chapter.

The goal of this thesis is to develop a solution to an additional identification problem that arises when fitting APC models to data that comes in unequal intervals which we called the curvature identification problem. To develop a model that resolves the curvature identification problem, we first describe the different forms of identification problems that are present in APC models and explain how they are resolved to motivate the solution we propose in later Chapters.

1.2.1 Overall level identification problem

First let's assume there is no structural link between the temporal covariates, that is $c \neq p - M \times a$. The overall level identification problem means we can add a term (in this case a constant) to each of the terms in the statistical model and not change the overall level. To see this, consider the transformed functions of age, period, and cohort

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

where const_1 and const_2 are constants. By considering the APC model in terms of the transformed linear functions, it is easy to verify the linear predictor is invariant to the addition of the constants

$$\tilde{\eta}_{apc} = [f_A(a) - \text{const}_1] + [f_P(p) + \text{const}_1 + \text{const}_2] + [f_C(c) - \text{const}_2] = \eta_{apc}.$$

Notice the indexing for the linear predictor is for all three temporal trends as there is no structural link between the three. This is different to Eq.(1.1) where the indexing is defined by only age and period.

The overall level identification problem is resolved by defining the APC model in terms of identifiable quantities. In this case, an identifiable quantity is one that is invariant to the addition of a constant. For example, the first derivatives (or first differences if using a discrete time factor

model) are invariant to the addition of a constant and

$$\tilde{f_A}'(a) = \frac{d}{da}\left[\tilde{f_A}(a) - \text{const}_1\right] = f'_A(a)$$

is identifiable. Similar expressions can be defined for period and cohort. Since the first derivatives are identifiable, the APC model can be reparameterised in terms of them to define an identifiable model.

The reparameterisation can be imposed through the well-known 'sum-to-zero' (also known as 'usual') constraints $\sum_a f_A(a) = \sum_p f_P(p) = \sum_c f_C(c) = 0$ (McCullagh and Nelder, 1989). Imposing these constraints on each covariate in the model is equivalent to defining the model in terms of the first derivatives which reduces the parameter space meaning it is no longer overparameterised.

An important implication of the reparameterisation is that the interpretation of the temporal covariates changes. Before sum-to-zero constraints, $f_A(a)$ is interpreted as the effect of age a on the response. After sum-to-zero constraints, $f_A(a)$ is interpreted as the effect of moving from one age to the next, it is a relative change.

1.2.2 Structural link

Now we return to the case where the structural link is present, i.e., $c = p - M \times a$. The structural link identification problem means the linear predictor is invariant to the addition of a linear trend. As with the overall level identification problem, to see this, consider the set of transformed temporal functions where we include the const_1 and const_2 terms since the structural link identification problem is compounded on-top of the overall level identification problem

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

As described previously, using $c = p - M \times a$ it is easy to confirm the linear predictor is invariant to the linear trend. In addition, the first derivatives are no longer identifiable since $\tilde{f_A}'(a) = f'_A(a) + \text{const}_3 M$ with similar for period and cohort.

With the temporal functions and their first derivatives invariant to the addition of a linear trend, the sum-to-zero constraints on their own are no longer appropriate to ensure identifiability in a model that considers all three temporal effects. We now consider two approaches that have been used to resolve the structural link identification problem. The first is using a constraint-based approach and the second is a reparameterisation based approach.

1.2.2.1 Constraint-based approach

A constraint-based approach aims to resolve the identification problem by forcing some of the parameters in the model to be equal. In doing this, there is no longer an overparameterisation as the constraint reduces the parameter space. The motivation behind why a particular constraint is chosen is in general due to one of two reasons: scenario specific reasoning or mathematical ease.

Examples of the use of a scenario specific constraint are numerous (Boyle et al., 1983; Robertson and Boyle, 1986; Boyle and Robertson, 1987) with the reasons and implementations. It is this variation of the chosen constraint that makes them unsuitable for use in general. They rely too heavily on both the scenario and the knowledge of the user. In contrast to methods generated with mathematical ease, making a solution based on scientific advice does not necessarily translate into a better overall model. Being able to provide a solution which is robust, applicable to different datasets without the need for expert backed constraints and is easily interpreted are qualities of a good model.

For any constraint-based approach, the general problem is the difficulty in understanding and interpreting the estimates. To explain this, we have fit five different APC models to an example data set in equal intervals. In each of the models, we resolved the structural link by constraining the effects of two sequential age groups to be equal. The APC estimates for each of the fits is shown in Figure 1-2. The noticeable difference between each of the fits shows how the constraint-based approach to resolving the structural link identification is extremely sensitive to the choice of constraint used. Since it is impossible to verify from the data alone which of the sets of estimates is indeed correct, solutions that use this form of constraint are generally not recommended.

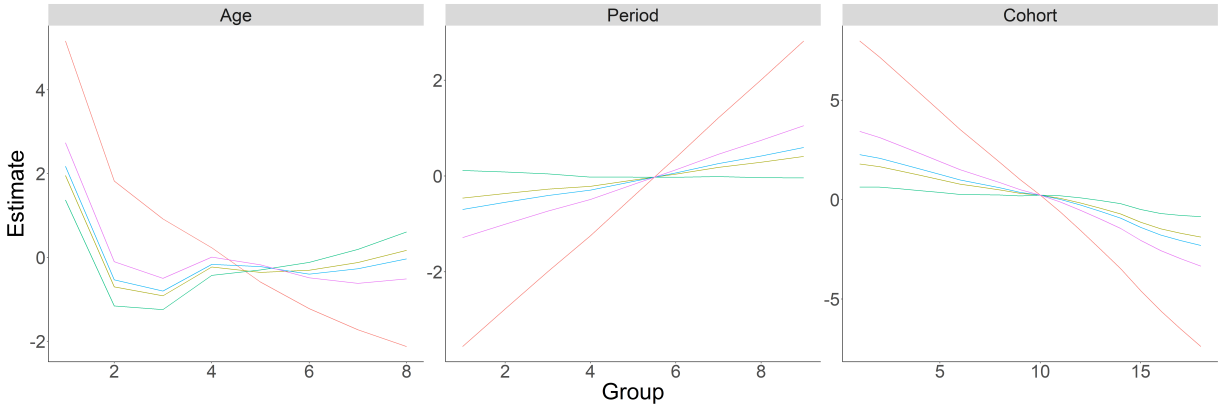


Figure 1-2: Estimates of age, period and cohort under sum-to-zero constraints and equality of different sequential pairs of age groups for data that is aggregated in equal intervals.

1.2.2.2 Reparameterisation based approach

In general, the best approaches follow in the footsteps of how a model is made identifiable for the overall level identification problem; it is reparameterised in terms of identifiable quantities. In the overall level identification, the identifiable functions are the first derivatives since they are invariant to a constant. Due to the structural link, the first derivatives are no longer identifiable but, in a similar vein, the second derivatives are invariant to both a constant and a linear trend,

$$\tilde{f_A}''(a) = \frac{d^2}{da^2}\left[\tilde{f}_A(a) + \text{const}_3 Ma\right] = f_A''(a)$$

with similar for period and cohort. The reparameterisation can be imposed using 'zero slope' constraints. An example of zero slopes constraints can be $\sum_a [a \times f_A(a)] = \sum_p [p \times f_P(p)] = \sum_c [c \times f_C(c)] = 0$, on each of the temporal terms alongside their sum-to-zero constraints.

Second derivatives (or an equivalent quantity) have been utilised by many authors to define an identifiable set of quantities to model all three temporal trends together. Holford (1983) reparameterises each temporal term in terms of a linear trend and a set of curvatures (these are equivalent to the second derivatives, i.e., $f_A'' \equiv f_{A_C}$) that are orthogonal to both the linear trend and a constant (intercept) term. Holford defines orthogonality with respect to the usual inner product $\langle x|y\rangle = \sum_i x_i y_i$ (Carstensen, 2007). Kuang et al. (2008) reparameterises each temporal term in terms of second differences (discrete time version of second derivatives) which are equivalent to the curvatures. Finally, Carstensen (2007) reparameterises the period and cohort terms in terms of curvatures like Holford but using a different definition of orthogonal (with respect to a weighted inner product), and then uses a reference period and cohort to centre them. Carstensen leaves age untouched saying this is the most important and should be a true age effect rather than a first difference.

Alongside the sets of identifiable quantities, the parameter space is completed with a few arbitrary chosen linear components. For example, Holford (1983) includes an intercept and two out of three linear slopes, Kuang et al. (2008) use three linearly independent first differences and Carstensen (2007) fixes a drift (temporal variation not described by period and cohort) parameter and reference period or cohort. In each case, the choice of the remaining terms is arbitrary, and we refer to the resulting model as 'overall non *ad-hoc*'.

1.2.2.3 Identifiable age-period-cohort model

Both approaches provide solutions to the structural link identification problem but since there is no way to check if a constraint is correct, constraint-based solutions should be considered with care. In contrast, the reparameterisation methods are preferred since they are based off identifiable quantities which are consistent. Following on from Holford (1983), we reparameterise

each temporal term in terms of a linear slope and a curvature function. The linear slope will be orthogonal to a constant (e.g., using sum-to-zero constraints) and the curvature will be orthogonal to a constant and its respective linear trend (e.g., using both sum-to-zero and zero-slope constraints).

The identifiable APC model we use for the most part of this thesis is,

$$\eta_{ap} = \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c) \tag{1.2}$$

where β_0 is the intercept, t_1 and t_2 are two of the three temporal slopes both orthogonal to a constant with parameters β_1 and β_2, respectively, and f_{A_C}, f_{P_C} and f_{C_C} are the age, period, and cohort curvature terms orthogonal to an intercept and their respective slope. The curvature functions are equivalent to second derivatives i.e., $f''_A \equiv f_{A_C}$ with similar for period and cohort.

As we mentioned previously, Holford (1983) completes the parameter space with an intercept and two out of the three linear slopes. To generalise what parameters are estimable in the curvature reparameterisation scheme, consider t_a, t_p and t_c, the respective age, period, and cohort linear trends. Whilst the individual slopes cannot be estimated, Holford showed that any linear combination of $\kappa_1 t_a + \kappa_2 t_b + (\kappa_2 - \kappa_1) t_c$ is estimable for arbitrary κ_1 and κ_2. In addition, since the curvatures are identifiable, the interpretation of them does not change depending upon the arbitrary choice of κ_1 and κ_2. An example choice of could be $\kappa_1 = 0$ and $\kappa_2 = 1$. In this case, the estimable parameter $t_p + t_c$ is known as the net-drift and cannot distinguish between period and cohort effects.

1.3 Unequal intervals

As has been described previously, it is common for data to come aggregated into unequal temporal intervals, and this could be due to several reasons such as: confidentiality, avoid zero counts, aid in the dissemination of the data, and simply, this is how it was collected. In comparison to the equal interval literature for APC models, the unequal interval literature is underdeveloped and sparse due to the curvature identification problem that is encountered when fitting APC models to data aggregated in unequal intervals.

Because of the curvature identification problem, users choose to collapse data into the simpler, but more restrictive case, of equal intervals data. To maximise the efficiency and make the most out of the data, we need methods that can appropriately deal with the curvature identification problem and handle data in any format.

When data comes in unequal intervals, on top of the overall level and structural link identification problems, there is the curvature identification problem. Consider the same toy data used

in Figure 1-2 but with the data aggregated into five-year age intervals and single-year period intervals. When aggregated like this, the ratio of the age-to-period aggregation is $M = 5/1$. Figure 1-3 shows the results of five sets of APC estimates found by constraining sequential pairs of age groups to resolve the structural link identification problem.

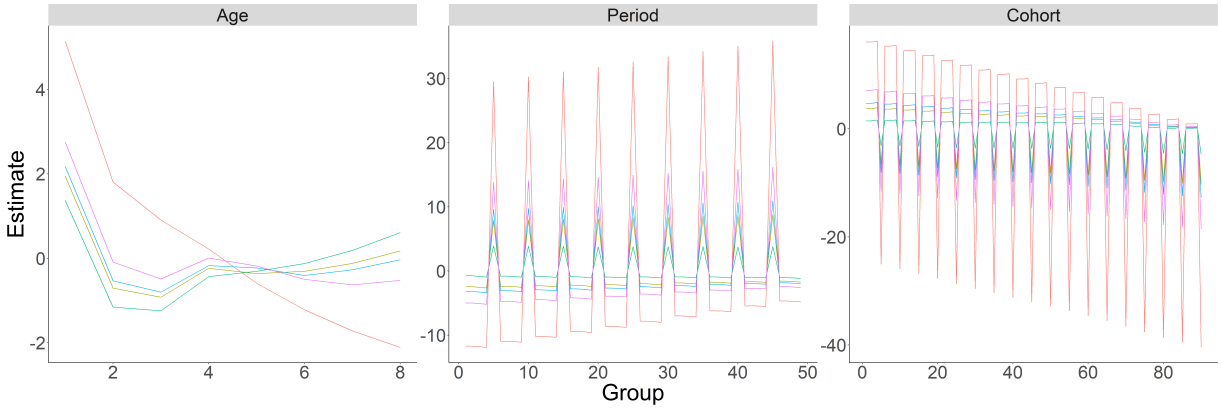


Figure 1-3: Estimates of age, period and cohort under sum-to-zero constraints and equality of different sequential pairs of age groups for data that is aggregated in unequal intervals.

Whilst the age estimates are unchanged from Figure 1-2, the period and cohort terms are displaying a cyclic (saw-tooth) pattern that is repeating every five period and cohorts, respectively, in each of the sets of estimates. Whilst we cannot say for definite if the cyclic pattern is an inherent property of the dataset, it seems unlikely that this is the case since this pattern was not present in each of the sets of period and cohort estimates when modelling the data in equal intervals. Therefore, if we assume the cyclic pattern is not in the data, we conclude it must be a symptom from fitting an APC model to unequal interval data, i.e., it is a symptom of the curvature identification problem. In addition, the cyclic pattern in each of the period and cohort estimates is repeating every five which matches the age-to-period aggregation ratio. This could be an indication the pattern is an artifact from the model fitting process. Similar conclusions have been reached by Holford (2006) and Held and Riebler (2012) when fitting APC models to unequal interval data.

As described previously, the curvature identification problem can be summarised in the following set of transformed functions which now include the cyclic function v_M to capture the cyclic pattern,

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

As before, the linear predictor is invariant to the addition of v_M due to $c = p - M \times a$ and $v_M(x + M) = v_M(x)$.

When data comes in equal intervals, we have shown that the second derivatives for all three temporal effects are identifiable. When data comes in unequal intervals, this is no longer the case,

$$\begin{aligned}\tilde{f_A}''(a) &= f_A''(a)\\ \tilde{f_P}''(p) &= f_P''(p) + v_M''(p)\\ \tilde{f_C}''(c) &= f_C''(c) - v_M''(c)\end{aligned}$$

due to the presence of the second derivative of the cyclic function v_M in the second derivative of the transformed period and cohort terms. As curvature is equivalent to the second derivatives, i.e., $f_A''(a) \equiv f_{A_C}(a)$, we can say the period and cohort curvature terms are unidentifiable when data comes in unequal intervals, hence why we have coined this the 'curvature identification problem'.

1.3.1 Previous methods for unequal interval data

To avoid the curvature identification problem altogether, the unequal interval data can be collapsed into equal interval data and in some cases, this can be beneficial but in others, it cannot. For example, if data comes in 2×1 and 3×5 formats, collapsing can be good and bad, respectively. In the former format, collapsing the single-year over two-years may lead to smoother results as there will be less noise. But in the latter format, the three-year interval will need to be collapsed over five groups and the five-year over three groups, resulting in a large amount of information being lost as well as greatly reducing the number of observations which will lead to an increase in the uncertainty of the estimates.

The choice of collapsing is a trade-off between information loss and ease of implementation. For the 2×1 example, the information loss is acceptable when compared against attempting to fit a more complicated model but in the 3×5 format, this is less so. As collapsing is an *ad-hoc* choice that is situation dependent, it does not transfer well to other scenarios and is rather restrictive. In addition, when attempting to conduct analysis on a given dataset, we want to extract the maximum amount of information out of the data and collapsing limits our ability to do this. It would be better to not have to make this choice in the first place and have a model that can deal with the data in whatever format we receive it in.

Another method is the use of continuous functions to represent the temporal effects. Since the curvature identification problem stems from how the data is tabulated, some authors have used, to a good success, continuous functions to model the temporal trends. Continuous functions do not rely on the tabulation of data and therefore, seemingly do not suffer from the curvature identification problem. When implementing models using continuous functions, one must always define a set of basis functions that gives a finite approximation of the continuous function to fit

a model practically.

One set of basis functions that are common in APC modelling are splines bases (Heuer, 1997; Holford, 2006; Carstensen, 2007), which are defined by a selection of points throughout the covariate called knots. Within the choice of which spline basis to use (for there are several different types), how they are defined (i.e., the number and placement of the knots) also must be considered. With this number of choices to make, the implementation of a spline has several different important decisions that need to be made. Often, these decisions are also scenario dependent.

As a final example, a method that has been used for fitting APC models to data in unequal intervals is to use smoothing priors within a Bayesian paradigm. For example, random walk (RW) priors are a common choice (Berzuini et al., 1993; Berzuini and Clayton, 1994; Besag et al., 1995; Knorr-Held and Rainer, 2001; Riebler and Held, 2010; Riebler et al., 2012a) for their properties with smoothing and low computational cost when implementing. In general, RW models penalise deviations from neighbours and the number of neighbours to consider (and hence, the definition of deviation) depends on the order of the RW. A RW of order two (RW2) model depends on two neighbours either side of the points in question and penalises deviations from a linear trend. RW priors do not explicitly tackle the curvature identification problem, instead focussing on smoothing over the cyclic pattern of the temporal estimates.

1.4 Smoothing

In this thesis, we fit smooth models in both a frequentist and a Bayesian paradigm. When fitting smooth models in the frequentist paradigm, we use smoothing splines and a penalised log-likelihood. In the Bayesian paradigm, we smooth using Gaussian Markov random fields (GMRFs). We now give an overview of how we smooth in both paradigms.

1.4.1 Frequentist smoothing using penalised log-likelihood

1.4.1.1 Penalised log-likelihood

In a frequentist paradigm, we consider the identifiable APC model to fall within the Generalized Additive Model (GAM) framework since the linear predictor depends on unknown smooth functions of our covariates (Hastie and Tibshirani, 1990). In general, a GAM has the form,

$$g(\mu_i) = \eta_i = \boldsymbol{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots$$

where $\boldsymbol{A}_i$ is a row of the model matrix for any strictly parametric terms (i.e., the intercept and two chosen temporal slopes), f_j are smooth functions of the covariates x_j (i.e., the smooth functions of age, period, and cohort curvature) and $i = 1, \ldots, n$.

Let $\mathcal{L}(\boldsymbol{f}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta})$ be the likelihood function for the set of observations. Estimates of the smooth function and other parameters in a GAM can be found by maximising the penalised log-likelihood with respect to $\boldsymbol{f}$ and $\boldsymbol{\theta}$,

$$\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\theta}} = \arg\max_{f,\theta} l(\boldsymbol{f}, \boldsymbol{\theta}) + \sum_j \lambda_j \int f_j''(x_j)^2 \, dx_j$$

where $l(\boldsymbol{f}, \boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{f}, \boldsymbol{\theta})$ is the log-likelihood and $\int f_j''(x_j)^2 \, dx_j$ is a penalty function for smooth f_j with smoothing parameter λ_j controlling the trade-off between model fit and smoothness.

The inclusion of the penalty function penalises the smooth functions when they deviate from linearity. Consequently, by smoothness we technically mean with respect to linearity as we consider straight lines to be the smoothest path between two points. If $\lambda_j = 0$, there is no cost for fitting complicated functions and $\hat{f}_j$ can be extremely 'wiggly'. As $\lambda_j \to \infty$, the cost for fitting a complicated function increases and $\hat{f}_j$ is forced to be closer to linearity.

1.4.1.2 Smoothing splines

In practise, the maximisation of the penalised log-likelihood needs the true functions f_j to have a finite basis. This is done by defining a finite approximation to the true function. In APC modelling, it is common to use a spline basis to approximate the true function f_j when maximising the penalised log-likelihood (Heuer, 1997; Holford, 2006; Carstensen, 2007). A spline basis is a set of polynomial (basis) functions which are based on points called knots. Given $b_k(x)$, the k^{th} basis function, f_j is approximated with a spline as follows

$$f_j(x_j) = \sum_{k=1}^{K_j} b_{jk}(x_j) \beta_{jk} = \boldsymbol{X}_j \boldsymbol{\beta}_j$$

where K_j is the number of basis function, β_{jk} are the unknown weights to be estimated and $\boldsymbol{X}_j$ is an $n \times K_j$ matrix of basis vectors. With this basis representation, the penalty function for f_j can be rewritten

$$\int f_j''(x_j)^2 \, dx_j = \boldsymbol{\beta}_j \int \boldsymbol{b}_j^T(x_j) \boldsymbol{b}_j(x_j) \, dx_j \boldsymbol{\beta}_j = \boldsymbol{\beta}_j \boldsymbol{S}_j \boldsymbol{\beta}_j$$

where $\boldsymbol{S}_j = \int \boldsymbol{b}_j^T(x_j) \boldsymbol{b}_j(x_j) \, dx_j$ is known as the penalty matrix. Therefore, the penalised log-likelihood to be maximised can be re-written in terms of the finite basis approximation to each

of the smooth functions,

$$\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}} = \arg\max_{\beta,\theta} l\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) + \sum_j \lambda_j \boldsymbol{\beta}_j \boldsymbol{S}_j \boldsymbol{\beta}_j.$$

There are several different spline bases one can use to represent a true function, and in this thesis, we use cubic regression splines when approximating the true function. In general, a regression spline is one that has a much smaller dimension than the data being analysed. The knots used to specify the basis functions of the splines should be arrange so that they cover the whole distribution of the covariate values in the original data set. The term cubic comes from the fact that the spline is piecewise cubic polynomial between the knots and is continuous up to the second derivatives (Wakefield, 2013). Cubic regression splines are computationally cheap and have directly interpretable parameters, but can only model one covariate at a time and the knots need to be predefined (Wood, 2017).

We fit penalised smoothing splines using `mgcv`, which can use a wide range of different bases for the approximation and estimation of a true function. In `mgcv` the penalised log-likelihood function is optimised using a penalised iterative re-weight least squares (PIRLS) approach (Wood, 2017). Whilst we choose to use a cubic regression spline, the methods we discuss are not limited solely to this spline basis.

1.4.2 Bayesian smoothing using Gaussian Markov random fields

1.4.2.1 Gaussian Markov random fields

We use Gaussian Markov random fields (GMRFs) for smoothing in a Bayesian paradigm. The following gives a brief overview of GMRFs, how they are used in smoothing and how we compute using them. We use the work of Rue and Held (2005), within which a more detailed exposition can be found. A GMRF is a random vector following a multivariate normal (Gaussian) distribution. For example, a vector $\boldsymbol{z} = (z_1, \ldots, z_n)$ is called a GMRF with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{Q}$ if its density is of the form

$$p\left(\boldsymbol{z}\right) = \frac{1}{\left(2\pi\right)^{n/2}} |\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{z} - \boldsymbol{\mu}\right)^T \boldsymbol{Q}\left(\boldsymbol{z} - \boldsymbol{\mu}\right)\right) \tag{1.3}$$

where $\boldsymbol{Q}$ is positive definite. The Markov property comes from the fact we require the GMRF to satisfy a conditional independence property,

$$z_i \perp z_j | \boldsymbol{z}_{-ij} \iff Q_{ij} = 0$$

where $\boldsymbol{z}_{-ij}$ are all elements in $\boldsymbol{z}$ without z_{ij} and $i \neq j$. Due to the conditional independence, most of the entries in the precision matrix $\boldsymbol{Q}$ are zero with only a few being non-zero, this means $\boldsymbol{Q}$ is a sparse matrix. Furthermore, we can represent the conditional independence using an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consisting of vertices $\mathcal{V}$ and edges $\mathcal{E}$. For temporal data, the graph $\mathcal{G}$ often appears as a line graph with $(i, i+1) \in \mathcal{E}$ for $i = 1, \ldots, n-1$ or $(i, i+2) \in \mathcal{E}$ for $i = 1, \ldots, n-2$ for the RW1 and RW2 models, respectively. Examples of these can be seen in Figures 1-4 and Figures 1-5, respectively. For spatial data, discussed in due course, $\mathcal{G}$ is known as the adjacency graph and an example can be seen in Figure 1-6.

Sparsity of a matrix is an important property as it offers excellent computational advantages over non-sparse (dense) matrices. Consider $\boldsymbol{Q}$ to be an $n \times n$ precision matrix. When $\boldsymbol{Q}$ is sparse, the order of the cost for any computations (such as inference) using $\boldsymbol{Q}$ is $\mathcal{O}\left(n^{3/2}\right)$. For the corresponding dense $\boldsymbol{Q}^{-1}$, the same computational costs are of order $\mathcal{O}\left(n^3\right)$ (Bakka et al., 2018).

1.4.2.2 Temporal smoothing

For APC modelling, the most common types of GMRFs that are used are RW1 and RW2 models (Berzuini et al., 1993; Berzuini and Clayton, 1994; Besag et al., 1995; Knorr-Held and Rainer, 2001; Riebler and Held, 2010; Riebler et al., 2012a). The popularity of both the RW1 and RW2 models in APC modelling is not only due to their smoothing properties with RW1 and RW2 models smoothing by penalising deviations in the constant and linear trends, respectively, but as they are GMRFs with sparse precision matrices, they offer good computational properties. The conditional independence structure of a RW can be shown through the undirected graph with Figure 1-4 showing $\mathcal{G}$ for a RW1 and Figure 1-5 showing $\mathcal{G}$ for a RW2. In both, the lines are the edges $\mathcal{E}$ and the nodes at which they meet are the vertices $\mathcal{V}$.

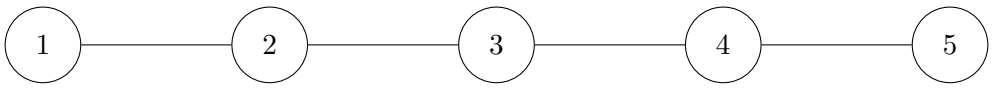


Figure 1-4: Graph for a RW1 model.

Since we wish to penalise deviations in linearity, we focus on the RW2 models using the work from Rue and Held (2005). Assuming $\boldsymbol{z}$ follows a RW2 model, the second differences have the distribution

$$\Delta^2 z_i \sim_{\text{iid}} \mathrm{N}\left(0, \tau^{-1}\right)$$

where τ is the precision parameter. The precision τ controls the trade-off between smoothness and closeness to the data (it is the smoothing parameter), and has parallels with the smoothing parameter λ in the GAM framework; for example, $\tau \to \infty$ shrinks the distribution towards the

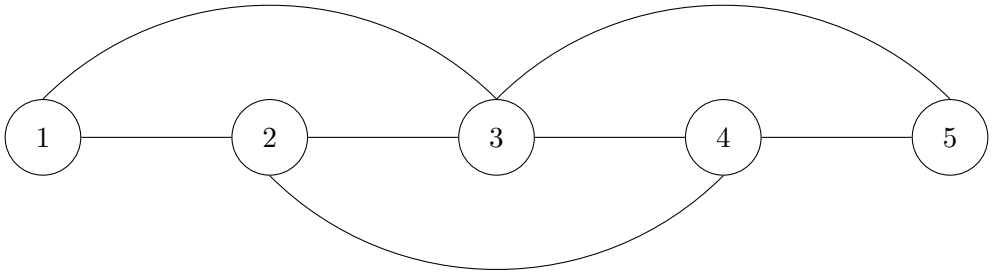


Figure 1-5: Graph for a RW2 model.

mean whilst $\lambda \to \infty$ smooths towards the mean. The joint density of $\boldsymbol{z}$ is therefore defined,

$$\begin{aligned} p(\boldsymbol{z}|\tau) &\propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n-2}(z_i - 2z_{i+1} + z_{i+2})^2\right) \\ &= \tau^{(n-2)/2} \exp\left(\frac{1}{2}\boldsymbol{z}^T\boldsymbol{Q}\boldsymbol{z}\right) \end{aligned} \tag{1.4}$$

which has the conditional mean and precision

$$\begin{aligned} \mathbb{E}(z_i|\boldsymbol{z}_i, \tau) &= \frac{4}{6}(z_{i+1} + z_{i-1}) - \frac{1}{6}(z_{i+2} + z_{i-2}) \\ \text{Prec}(z_i|\boldsymbol{z}_i, \tau) &= 6\tau \end{aligned}$$

respectively for $2 < i < n-2$. The precision matrix for a RW2 has the form

$$\boldsymbol{Q} = \tau\boldsymbol{R} = \tau \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}$$

where $\boldsymbol{R}$ is referred to as the structure matrix. Due to the Markov property, the RW2 precision matrix is of rank $n-2$ and is rank deficient. Therefore, Eq.(1.4) is technically not a proper distribution meaning it is no longer a GMRF, but an improper GMRF. In general, a RW of order p model is an improper GMRF with the precision matrix being of rank $n-p$. An improper GMRF cannot be sampled from, but as will be described in Chapter 3, an improper GMRF is equivalent to a GMRF on a lower density; therefore, under appropriate constraints we can sample from the improper GMRFs.

1.4.2.3 Spatial smoothing

We have described in detail how temporal data is important for measuring health related outcomes. Another important type of data is spatial data. Spatial data arise when the outcome of interest is measured at points (e.g., regions) within the domain of interest. For example, incidence of a disease across a country with the incidence measured at each region within the country. In health-related outcomes, it is important to understand the geographical pattern. For example, we believe that areas that are closer together are more likely to share certain characteristics that influence a given health outcome than regions further apart. Therefore, the regions closer together will have a similar response than those further apart.

Alongside the temporal scales, it is common for most providers of health and demographic data to include additional covariates of interest. For example, both the ONS weekly all-cause mortality and the DHS survey data contain a covariate for the region the observation was measured in. Therefore, extending the identifiable APC model we have proposed to include a spatial and/or spatio-temporal (an interaction between a spatial and a temporal component) term is natural. No further considerations need to be made relating to the identifiability of the APC terms alongside the spatial components if the APC terms are identifiable (Smith, 2018).

Whilst it is possible to perform spatial modelling in the frequentist paradigm, we choose to model the spatial structure in a Bayesian paradigm to exploit the relationship between the spatial structure and GMRFs for computational gain. This is a crucial aspect for spatial modelling since most spatial models often involve a complex model specification and large datasets. To make this exploitation, we first define a neighbourhood across the domain of interest. Two regions are called neighbours if they share a common border. This network of neighbours can be considered as an adjacency graph $\mathcal{G}$. For example, Figure 1-6 is a map of Kenya with the adjacency graph $\mathcal{G}$ superimposed. The lines are the edges $\mathcal{E}$ and the points at which they meet are the vertices $\mathcal{V}$.

The neighbourhood, represented by $\mathcal{G}$, has a conditional independence property. Thus, we can use the GMRF framework to model the spatial covariate. To see this, let $\mathcal{N}(i)$ represent the neighbourhood region of i then for $i \neq j$,

$$z_i \perp z_j | \boldsymbol{z}_{-ij} \iff Q_{ij} = 0 \iff j \notin \mathcal{N}(i)$$

where $\boldsymbol{z}$ a GMRF with respect to the graph $\mathcal{G}$ (Rue and Held, 2005). In other words, if z_i and z_j are conditionally independent, the corresponding element in the precision matrix is zero and i and j are not neighbours. Using the conditional independence between neighbours, we can derive a (generally sparse) precision matrix from the Kenyan adjacency graph. Let $i \sim j$ represent when i and j are neighbours, Figure 1-7 shows the precision matrix corresponding to Kenya where the black squares represent both $i \sim j$ and $i = j$ and the white spaces represent zeros as i and j are not neighbours.

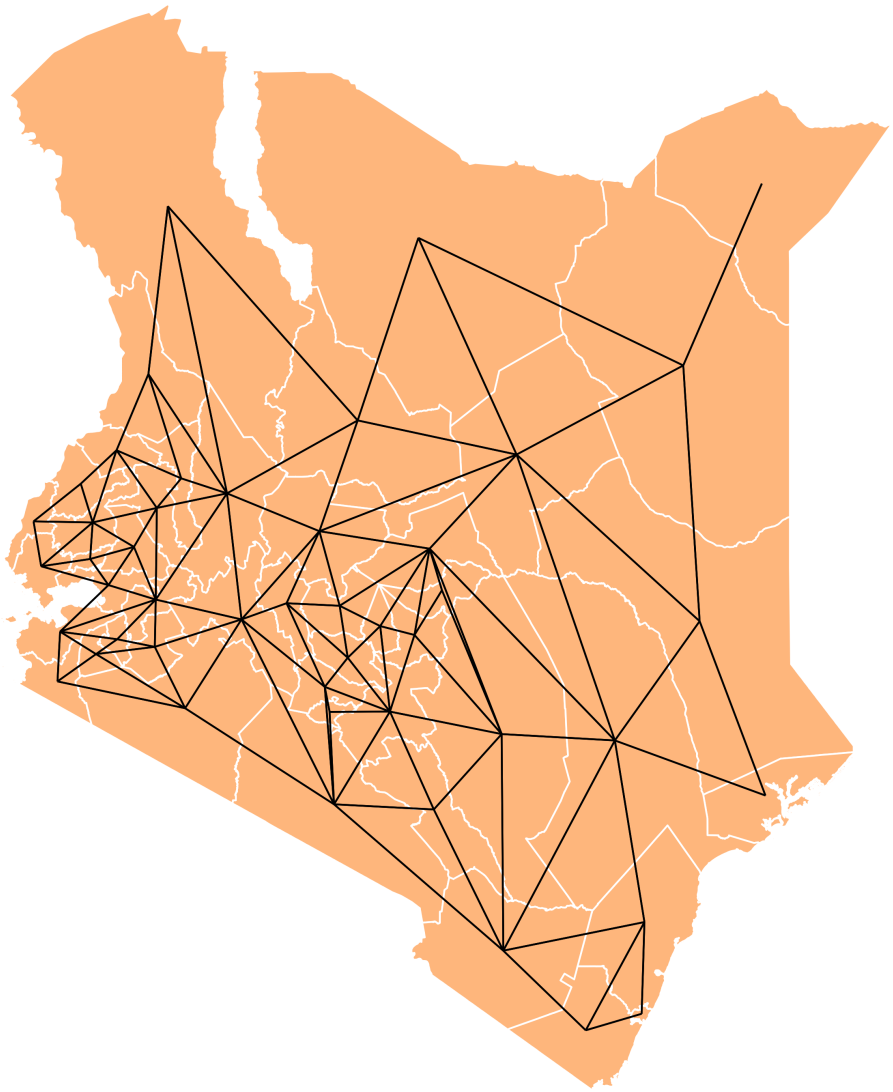


Figure 1-6: Adjacency spatial polygon for Kenya. The black vectors indicate the neighbouring regions.

For spatial modelling, a commonly used GMRF to capture the spatial dependence is an intrinsic conditional autoregressive (ICAR) model Besag et al. (1991). Assume $\boldsymbol{z}$ follows a ICAR model,

the joint density is defined by

$$p(\boldsymbol{z}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{i\sim j}(z_i - z_j)^2\right)$$
$$= \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{i\sim j}\boldsymbol{z}^T\boldsymbol{Q}\boldsymbol{z}\right)$$

which has the conditional mean and precision

$$\mathbb{E}(z_i|\boldsymbol{z}_i, \tau) = \frac{1}{\mathcal{N}_i}\sum_{j:i\sim j} z_j$$
$$\text{Prec}(z_i|\boldsymbol{z}_i, \tau) = \mathcal{N}_i\tau.$$

where $\mathcal{N}_i$ is the number of regions in the neighbourhood of i and the precision matrix of an ICAR has the form

$$Q_{ij} = \tau \begin{cases} \mathcal{N}_i & \text{if } i = j \\ 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

(Rue and Held, 2005). The term "intrinsic" is due to the precision matrix $\boldsymbol{Q}$ for the ICAR model being rank deficient; therefore, the ICAR model is an improper GMRF.

Consider the spatial term $\boldsymbol{z}$ partitioned into structured, $\boldsymbol{u}$, and unstructured, $\boldsymbol{v}$, components such that,

$$\boldsymbol{z} = \boldsymbol{u} + \boldsymbol{v}$$

where $\boldsymbol{u} \sim \text{N}\left(0, \tau_u^{-1}\boldsymbol{Q}\right)$ is modelled by the ICAR model and $\boldsymbol{v} \sim \text{N}\left(0, \tau_v^{-1}\boldsymbol{I}\right)$ accounts for random spatial effects not captured by the structured component (overdispersion). This is known as the Besag-York-Mollié (BYM) model (Besag et al., 1991). In the BYM model, the unstructured term cannot be seen fully independently from the structured term; therefore, they suffer from a lack of identifiability (Eberly and Carlin, 2000; MacNab, 2011). Due to this, care should be taken with the prior selection for τ_u and τ_v to ensure they are dependent on one another (Simpson et al., 2017). Additionally, as the precision matrix is dependent on the adjacency graph, interpretations of the precision parameters change if the graph is changed. This lack of scaling in the parameters between applications is of concern, since it does not allow comparisons of the parameters (e.g., precision, variance) across different spatial structures (Sørbye and Rue, 2014).

To account for these issues, using the recommendation of Simpson et al. (2017) to use a scaled structured spatial component $\boldsymbol{u}_\star$ with a scaled precision matrix $\boldsymbol{Q}_\star$, Riebler et al. (2016) define the so-called BYM2 model

$$\boldsymbol{z} = \frac{1}{\tau_z}\left(\sqrt{\phi}\boldsymbol{u}_\star + \sqrt{1-\phi}\boldsymbol{v}\right)$$

which has the covariance $\text{Var}(\boldsymbol{z}|\tau_b, \phi) = \tau_b^{-1}\left[(1-\phi)\boldsymbol{I} + \phi\boldsymbol{Q}_\star^{-1}\right]$. In the BYM2 model, $\phi \in [0,1]$ is known as the mixing parameter that attributes how much the marginal variance is explained by the structured spatial component. When $\phi = 0$, the model reduces to pure overdispersion and when $\phi = 1$, the model reduces to the ICAR model. Additionally, under the new parameterisation, the terms τ_b and ϕ are identifiable and the variance, $\text{Var}(\boldsymbol{z}|\tau_b, \phi)$, has the same interpretation across different graphs.

In Chapter 4 we extend the identifiable APC model, Eq. (1.2), to include structured and unstructured spatial covariates and a space-time covariate. By space-time we technically mean space-period, but the naming convention in the literature is space-time. The spatial covariates measure the influence a geographical point (such as a region) has on the health outcome and the space-time covariate measures the interaction between the spatial location and the period, allowing for a yearly variation within each spatial location considered. For $r = 1, \ldots, R$ a set of regions, an identifiable spatio-temporal APC model is,

$$\eta_{aps} = \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c) + u(r) + v(r) + \delta(p, r)$$

where $u(r)$ and $v(r)$ are the structured and unstructured spatial functions and $\delta(p, r)$ is the space-time interaction function. In the application to U5MR, we reparameterise the structured and unstructured spatial terms and model them together using a BYM2 model. The temporal and spatial components of the space-time term are modelled using the RW2 and ICAR model, respectively.

1.4.2.4 Computation

In a Bayesian paradigm, we consider both the APC and spatio-temporal APC model to fall within a LGM framework. A LGM is of the form,

$$g(\mu_i) = \eta_i = \boldsymbol{A}_i\boldsymbol{\beta} + \sum_j f_j(x_j)$$

where $\boldsymbol{A}_i$ is a row of the model matrix for any strictly parameteric terms and f_j are the non-linear smooth functions of the covariates x_j (Rue et al., 2009). This is like a GAM but without the multi-dimensional functions. Inference is obtained using a three-staged Bayesian hierarchical model (BHM) framework where the parameters are nested within one another. In this form, we have a flexible model to perform inference on different levels of the model such as the underlying latent process (Gelman et al., 2013).

In order to describe the BHM, gather all the model parameters in the linear predictor into a latent field $\boldsymbol{z} = \{\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{f}\}$ and let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ be the hyperparameters for the likelihood of the data $\boldsymbol{y}$ and prior of the latent field $\boldsymbol{z}$. In the BHM, the observations $\boldsymbol{y}$ are assumed to be conditionally

independent, given the latent Gaussian random field $\boldsymbol{z}$ and the hyperparameters $\boldsymbol{\theta}$. The three stages themselves are: the likelihood model, the latent model, and the hyperparameters. More explicitly this is

$$\begin{aligned} \text{Likelihood:} \quad & \boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}_1 \sim \prod_i p\left(y_i|z_i, \boldsymbol{\theta}_1\right) \\ \text{Latent Gaussian Field:} \quad & \boldsymbol{z}|\boldsymbol{\theta}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}_2\right), \mathbf{Q}\left(\boldsymbol{\theta}_2\right)^{-1}\right) \\ \text{Hyperparameters:} \quad & \boldsymbol{\theta} \sim p\left(\boldsymbol{\theta}\right) \end{aligned}$$

Where $\boldsymbol{\mu}\left(\boldsymbol{\theta}_2\right)$ and $\boldsymbol{Q}\left(\boldsymbol{\theta}_2\right)$ are the mean and precision of $\boldsymbol{z}$. In the RW2 and ICAR models, $\boldsymbol{\mu}\left(\boldsymbol{\theta}_2\right) = 0$.

For computation, we chose to use the integrate nested Laplace approximation (INLA) approach of Rue et al. (2009). Rather than approximating the full posterior distribution, INLA approximates the marginal posterior distribution by making a series of Laplace approximation and numerical integrations. In addition to not sampling from the full posterior, the INLA approach makes exceptional use of the structure of the precision matrix $\boldsymbol{Q}$, which for a GMRF is in general sparse due to the conditional independence property. During the numerical integrations, the sparse nature of $\boldsymbol{Q}$ is an essential property to ensure this process is computationally efficient. A sparse $\boldsymbol{Q}$ can still be utilised by traditional Markov chain Monte Carlo sampling methods to approximate the full posterior distribution. However, the approximation of the full posterior is less efficient than the approximation of the marginal posterior.

1.5 Thesis structure

The structure of the remainder of this thesis is as follows:

- In Chapter 2, using penalised smoothing splines, we show why a penalty function is critical to alleviating the curvature identification problem
- In Chapter 3, we consider the similarities between penalised smoothing spline and random walk prior models and the random walk prior models' ability to resolve the curvature identification problem
- In Chapter 4, we provide a detailed application of our proposed age-period-cohort model to find subnational estimates of under-five mortality rates in Kenya
- In Chapter 5, we provide a summary of the results and contributions of the thesis as presented in Chapter 2 through to Chapter 4. Conclusions and potential future opportunities for age-period-cohort models are discussed. Finally, the key impacts of the thesis are summarised

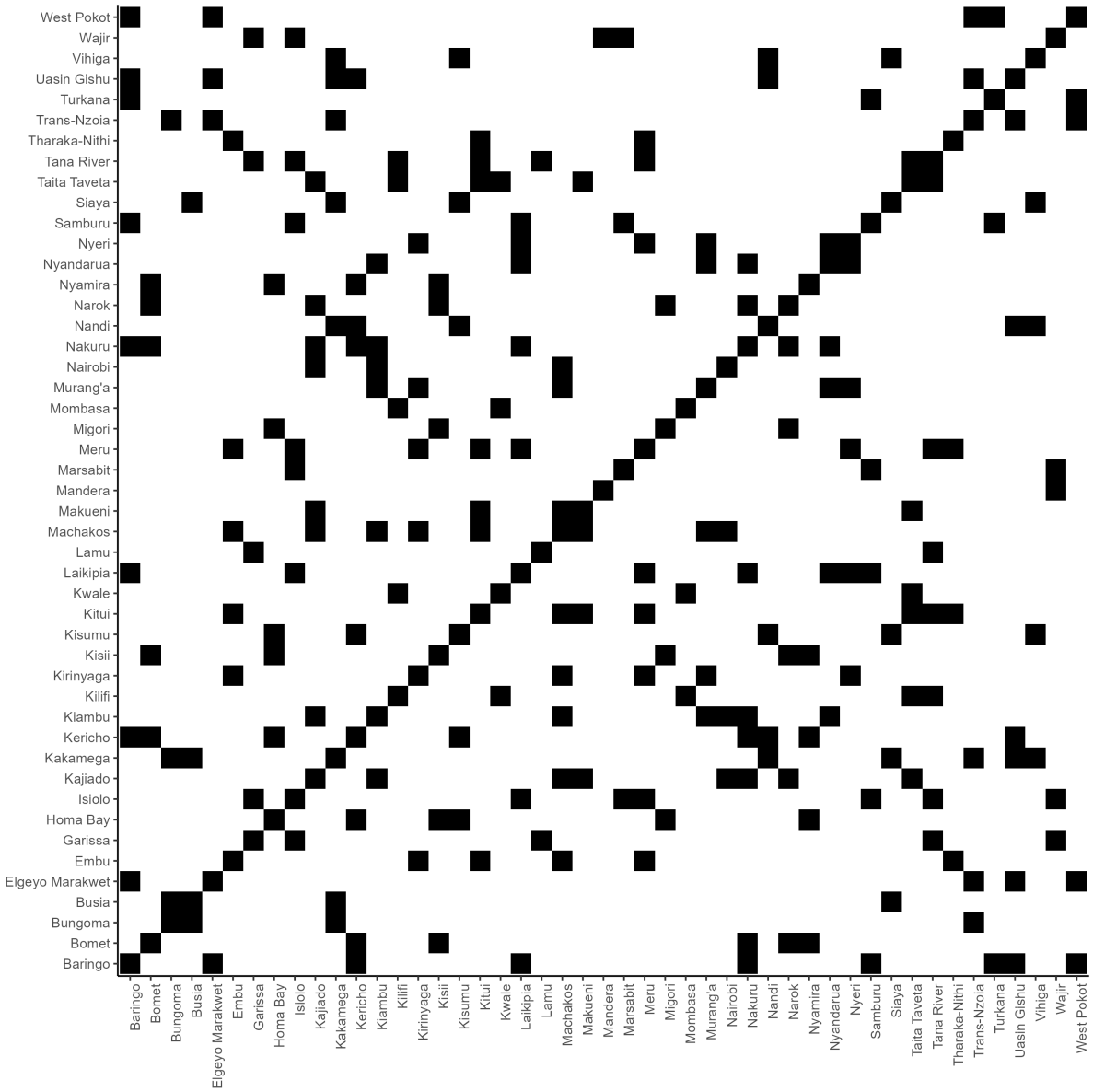


Figure 1-7: Precision matrix for Kenya. The black blocks indicate two regions that are neighbours and the white blocks indicate two regions that are not neighbours.

CHAPTER 2

PENALISED SMOOTHING SPLINES RESOLVE THE CURVATURE IDENTIFICATION PROBLEM IN AGE-PERIOD-COHORT MODELS WITH UNEQUAL INTERVALS

This chapter is reproduced from the authors manuscript that is currently undergoing revisions based on peer-review comments from Statistics in Medicine. A preprint can be found on arXiv (Gascoigne and Smith, 2021). Any changes between this and the version on arXiv are based on reviewer comments and styling to fit in with the rest of the thesis.

Using penalised smoothing splines, we demonstrate how the inclusion of a penalty is critical to providing an appropriate solution to the curvature identification problem; whereas, the use of factor models and un-penalised smoothing spline models do not. We do this using both theoretical and empirical results.

We propose a novel use, specification, and implementation of a penalty function to alleviate the curvature identification problem. We give a thorough explanation of how to reparameterise both the smoothing spline and the accompanying penalty to ensure that both are orthogonal to an intercept and linear trend as well as the basis functions being penalised are exactly those that capture the temporal curvature. To ensure accessibility for a wider audience, we describe how to implement this change in software that is commonly used for fitting penalised splines. A sensitivity analysis on the specification of the basis function was used to highlight how an un-penalised smoothing spline does not provide a set of estimates that are robust to the specification of the basis and the alleviation of the curvature identification problem; whereas, the penalised smoothing spline does. This leads to the conclusion that a penalty is a necessary inclusion when fitting APC models to data in unequal intervals.

Abstract

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).

2.1 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 (NHS, 2020). 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 (Holford, 1983) (terms that are orthogonal to a linear term) for each temporal effect, whilst Kuang, Nielson and Nielson based a solution on estimable second differences (Kuang et al., 2008) (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 (Carstensen, 2007). Smith and Wakefield offer a comprehensive review on APC models for data aggregated in equal intervals (Smith and Wakefield, 2016).

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 (ONS, 2020a) and, for a finer understanding of seasonality, weekly periods, and five-year age groups (ONS, 2020b). In addition, the Demographic and Health Surveys (DHS) release data for monthly ages and yearly periods

(USAID, 2019). 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 (Holford, 2006). Figure 2-1 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.

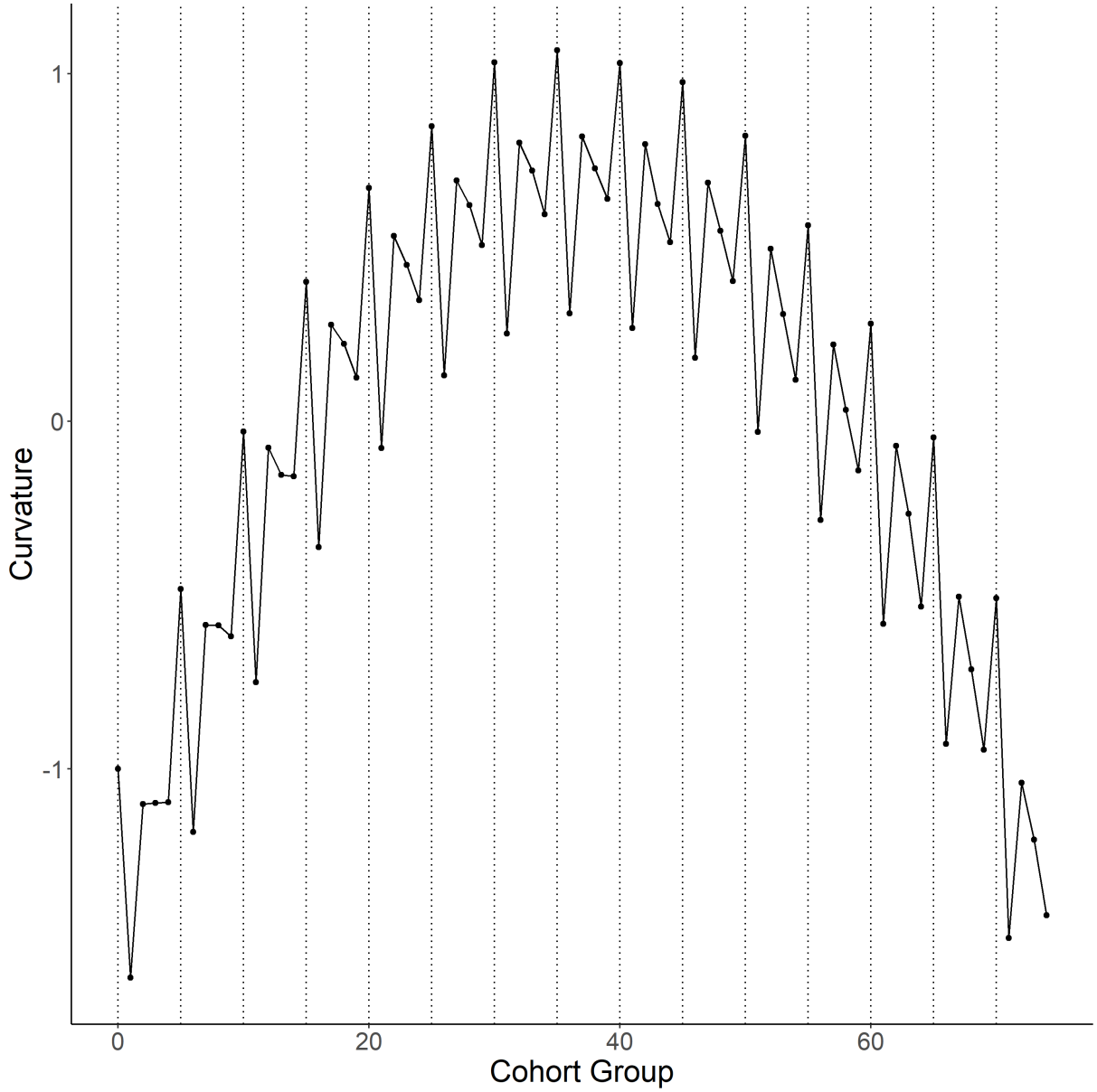


Figure 2-1: Cohort curvature estimate from fitting an APC model reparameterised into linear terms and their orthogonal curvatures to simulated unequal interval APC data.

As Figure 2-1 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 (Holford, 2006; Carstensen, 2007). 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 (Held and Riebler, 2012), 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 (Heuer, 1997), and Carstensen promotes the use of continuous functions in his reparameterisation (Carstensen, 2007). 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 chapter 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.

2.2 Method

2.2.1 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 2.1 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 2.1: Cohort indexing for age-period data table where age is grouped $M = 1$ times larger than period. The cohort index is defined using $c = M \times (A - 1) + P$ 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 $a = 1, 2, \ldots, A$ and $p = 1, 2, \ldots, P$. A cohort index is not explicitly defined due to the structural link, but is calculated $c = 1, 2, \ldots, C = M \times (A - a) + p$. A continuous APC model is

$$g(\mu_{ap}) = f_A(a) + f_P(p) + f_C(c) \tag{2.1}$$

where $g(\cdot)$ is the link function, $\mu_{ap} \equiv \mathbb{E}[y_{ap}]$ 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 (Carstensen, 2007)

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

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

$$\begin{aligned}&g(\tilde{\mu}_{ap}) \\ &= \tilde{f_A}(a) + \tilde{f_P}(p) + \tilde{f_C}(c) \\ &= [f_A(a) - \text{const}_1 + \text{const}_3 Ma] + \\ &\quad [f_P(p) + \text{const}_1 + \text{const}_2 - \text{const}_3 p] + \\ &\quad [f_C(c) - \text{const}_2 + \text{const}_3 c] \\ &= [f_A(a) + \text{const}_3 Ma] + [f_P(p) - \text{const}_3 p] + [f_C(c) + \text{const}_3 c] \\ &= f_A(a) + f_P(p) + f_C(c) \\ &= g(\mu_{ap})\end{aligned}$$

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

Many reparameterisation schemes are based off of an identifiable set of quantities. The first derivatives are not identifiable (Mason et al., 1973; Fienberg and Mason, 1979) but the second derivatives, or more generally curvatures, are identifiable. The age curvature is expressed

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

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

2.2.2 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

$$g(\mu_a) = f_A(a) \tag{2.2}$$

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` (Carstensen, 2007) package in `R` (R Core Team, 2021) 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 (Rutherford et al., 2010).

To approximate f_A in Eq.(2.2), 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 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 $\hat{\beta}_i$ (Wood, 2017). PIRLS is used to find an estimate $\hat{\boldsymbol{f}}_A$ that minimises the objective function

$$D(f_A(a)) + \lambda_A \int f''_A(a)^2\, da$$

where $D(f_A(a))$ 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). By representing the smooth function via a spline basis, the smooth itself can be written

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

for $\boldsymbol{X}$ an $n \times I$ matrix and $\boldsymbol{\beta}$ an $I \times 1$ 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.

2.2.3 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 $\mathbf{1}^T \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}$, avoiding any intercept related identification problems. The constraint is applied by constructing an $I \times (I-1)$ matrix $\boldsymbol{Z}$ through the QR-decomposition of $\left(\mathbf{1}^T \boldsymbol{X}\right)^T$. The smooth is reparameterised by using $\boldsymbol{X}\boldsymbol{Z}$ and $\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}$ as its model and penalty matrices; for more details, see Chapter 5 (Wood, 2017).

The parameter space of f_A can be split further into a linear slope and parameters corresponding to orthogonal curvatures (Holford, 1983). 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 (Carstensen, 2007). In the same vein as the intercept reparameterisation, define a $2 \times I$ array consisting of a constant and vector of all ages for the intercept and linear terms, $[\mathbf{1} : \boldsymbol{a}]$. Consequently, a $(I-1) \times (I-2)$ matrix $\boldsymbol{Z}$ is calculated by the QR-decomposition of $\left([\mathbf{1} : \boldsymbol{a}]^T \boldsymbol{X}\right)^T$, and the smooth f_A is reparameterised using $\boldsymbol{A}_C = \boldsymbol{X}\boldsymbol{Z}$ and $\boldsymbol{S}_{A_C} = \boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{Z}$ as its model and penalty matrices.

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

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

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,

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

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 $\lambda_A \boldsymbol{\beta}_{A_C}^T \boldsymbol{S}_{A_C} \boldsymbol{\beta}_{A_C}$.

Figure 2-2 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

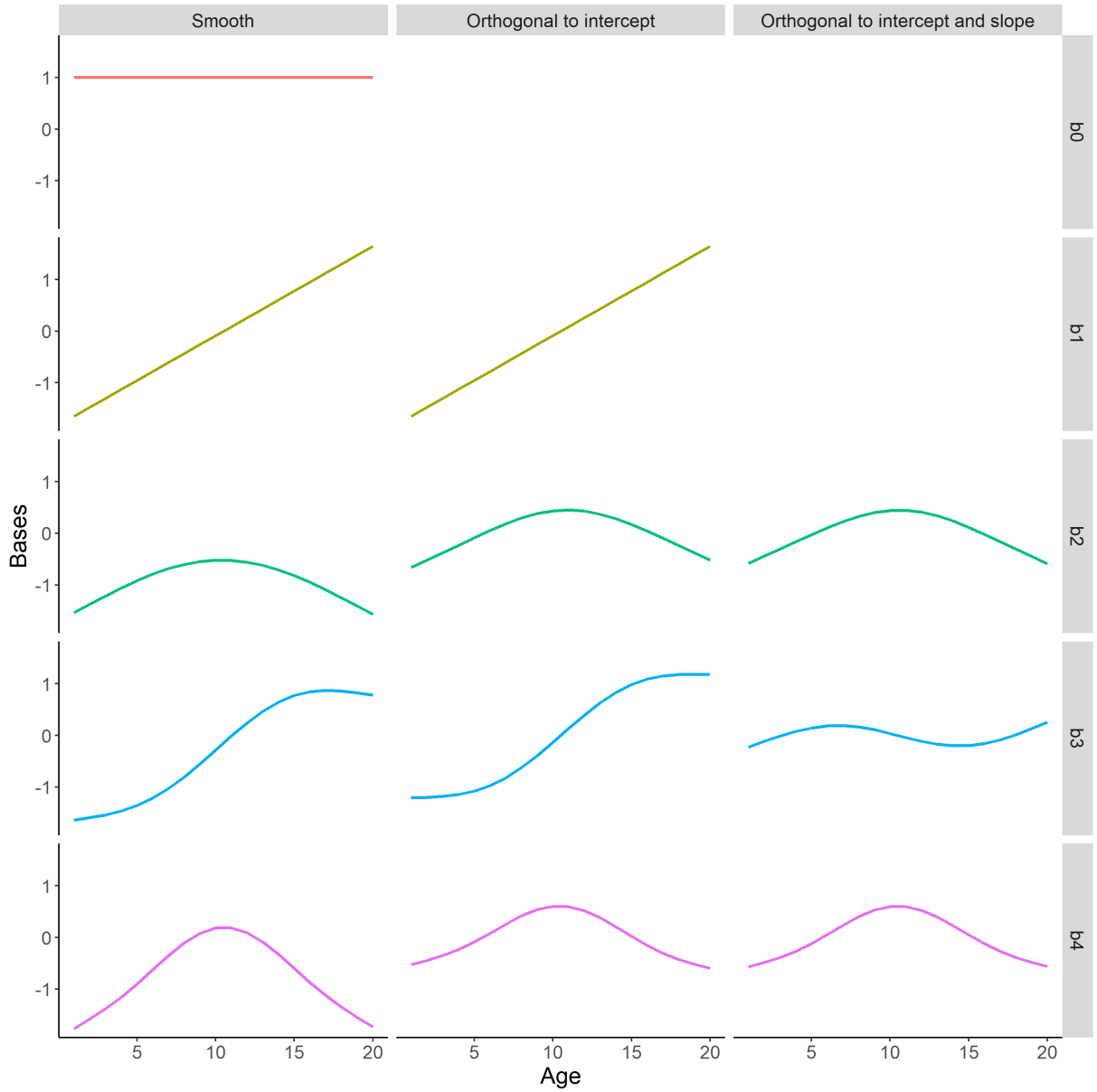


Figure 2-2: Thin plate regression spline basis before any reparameterisation, after an intercept reparameterisation and after an intercept and slope reparameterisation.

trends are removed. This is shown in Figure 2-2 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.

2.2.4 Age-period-cohort modelling

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

$$g\left(\mu_{ap}\right) = \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}\left(a\right) + f_{P_C}\left(p\right) + f_{C_C}\left(c\right) \tag{2.3}$$

where t_1 and t_2 are two of the three temporal slopes with parameters β_1 and β_2, and $f_{A_C}\left(a\right)$, $f_{P_C}\left(p\right)$ and $f_{C_C}\left(c\right)$ 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 (Holford, 1983). 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 t_a, t_p and t_c be the respective age, period, and cohort linear terms. Any linear combination of $\kappa_1 t_a + \kappa_2 t_p + \left(\kappa_2 - \kappa_1\right) t_c$ is estimable for arbitrary κ_1 and κ_2 (Holford, 1983). 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 (Holford, 1983), second differences (Kuang et al., 2008) (second differences are the discrete way to define local curvature, like the continuous second derivatives) and period and cohort terms orthogonal to linear trends (Carstensen, 2007) provide systematic solutions to the APC problem. In each scheme, an arbitrary choice is made: which linear slope to drop (Holford, 1983), which three baseline rates to choose (Kuang et al., 2008) and which term and reference term to use (Carstensen, 2007). 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.,

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

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

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

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.

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,

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

2.2.5 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). In Figure 2-2, 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.(2.3) 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) (Hastie and Tibshirani, 1990). We implement the GAMs in the `mgcv` package (Wood, 2017) in `R` (R Core Team, 2021) 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 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.

2.3 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.

2.3.1 Data

The simulation study is motivated by obesity rates and how they increase with age and in more recent years (Flegal et al., 2002; Mokdad et al., 2003; Ogden et al., 2006), as well as having an hereditary effect (Cole et al., 2008; Kuh and Shlomo, 2004). 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 (Luo and Hodges, 2016). 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) (Fannon et al., 2021). 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 (Luo and Hodges, 2016). 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 (NHS, 2020). 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

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

to match the obesity survey data. The expected responses for the binomial (overweight, BMI $\geq$ 25) and Poisson (obese, BMI $\geq$ 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

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

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

In this chapter, we will only report on the results for binomial generated data. The results for the other distributions are in Appendix A. Furthermore, Appendix A 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 reparameterisation we propose.

2.3.2 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 (Holford, 1983).

The factor version of the reparameterised APC model is,

$$g(\mu_{ap}) = \beta_0 + t_1\beta_1 + t_2\beta_2 + \alpha_{C_a} + \tau_{C_p} + \gamma_{C_p}$$

where t_1 and t_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[1] from the National Cancer Institute (Rosenberg et al., 2014).

2. Smoothing spline models: Detailed in Section 2.2, a reparameterisation in the style of the FA model but on a continuous version of the APC model using smoothing splines on the curvatures

$$g(\mu_{ap}) = \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c)$$

where t_1 and t_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 2.4.

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.

2.3.3 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

[1] https://analysistools.cancer.gov/apc/

marginal temporal effect of the linear predictor. For example, the true adjusted age effect for all three distributions are calculated,

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

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` $= 1$,

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

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

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

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, $\hat{\mu}_{apc}$, $\text{logit}(\hat{\pi}_{apc})$ 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 2-3. 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 are 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.

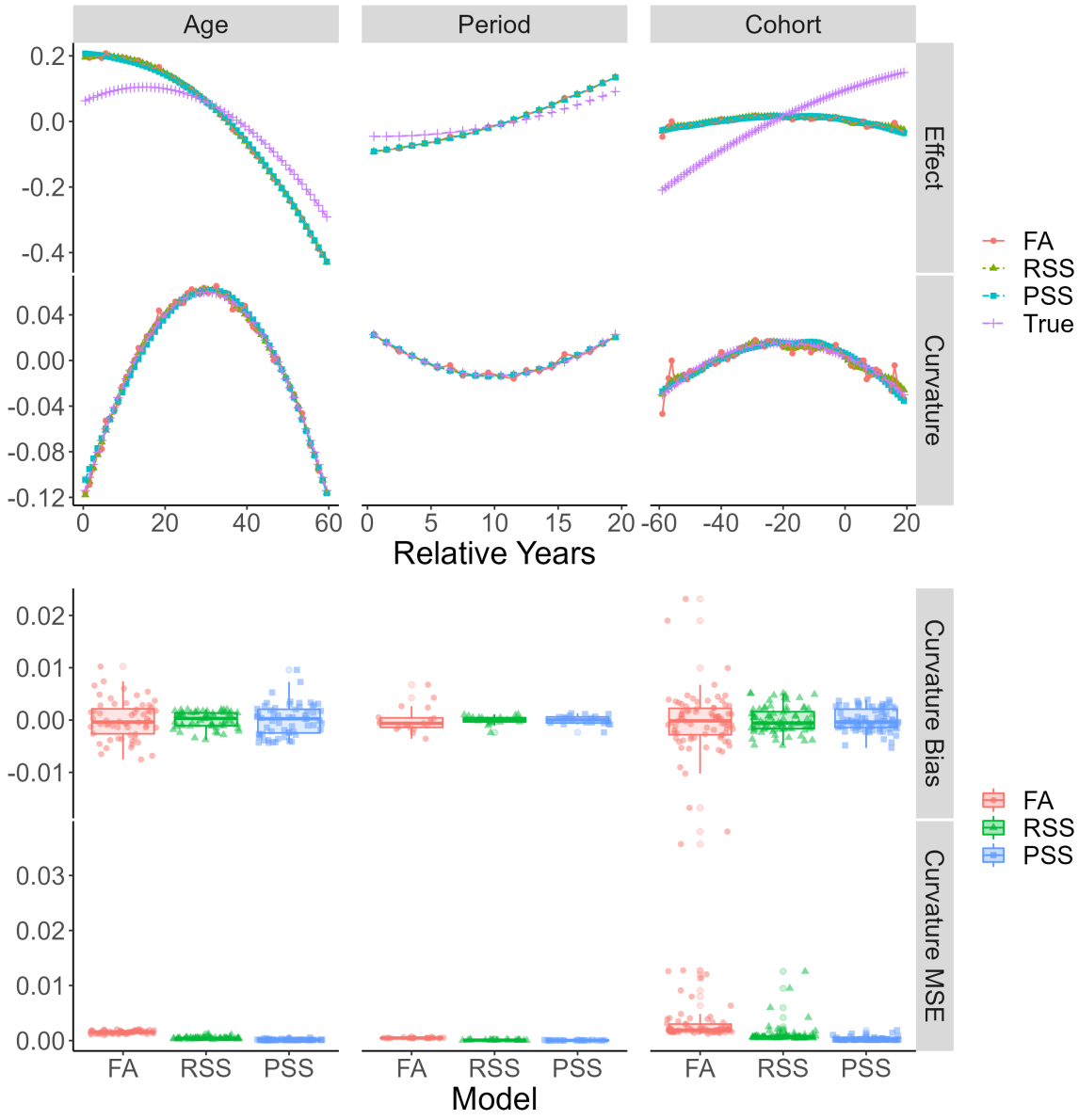


Figure 2-3: Simulation study results for equal interval, $M = 1$, binomial data generated when all three temporal effects are present. The FA, RSS and PSS models are the factor, regression smoothing spline and penalised smoothing spline models, respectfully. The first and second row are of the temporal 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.

The first row in Figure 2-3 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 (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 A-4 in Appendix A. Figure A-4 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.

In contrast to the full effect, the temporal curvatures are identifiable, and this is shown in the second row of Figure 2-3 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.

2.4 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.

2.4.1 Curvature identification problems

Previously we have focused on the case where age and period are in equal intervals, $M = 1$. Table 2.2 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 2.1, where cohorts appear every period.

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									

Table 2.2: 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.

As well as the identification problems for equal intervals, the additional identification issues that arise when modelling unequally aggregated APC data, $M \neq 1$, are encapsulated by the inclusion of two M periodic functions $v_M(p)$ and $v_M(c)$ in a set of transformed functions. 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

$$\begin{aligned} \tilde{f_A}(a) &= f_A(a) - \text{const}_1 + \text{const}_3 M a, \\ \tilde{f_P}(p) &= f_P(p) + v_M(p) + \text{const}_1 + \text{const}_2 - \text{const}_3 p, \\ \tilde{f_C}(c) &= f_C(c) - v_M(c) - \text{const}_2 + \text{const}_3 c. \end{aligned} \tag{2.4}$$

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

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

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

Now define the second derivatives of the two periodic functions as

$$\begin{aligned}
v''_M\left(p\right) &\equiv v_{M_C}\left(p\right) \\
v''_M\left(c\right) &\equiv v_{M_C}\left(c\right).
\end{aligned}$$

The previously identifiable second derivatives

$$\begin{aligned}
\tilde{f_A}''\left(a\right) \equiv \tilde{f_{A_C}}\left(a\right) &= f_{A_C}\left(a\right) \\
\tilde{f_{P_C}}\left(p\right) &= f_{P_C}\left(p\right) + v_{M_C}\left(p\right) \\
\tilde{f_{C_C}}\left(c\right) &= f_{C_C}\left(c\right) - v_{M_C}\left(c\right)
\end{aligned} \tag{2.5}$$

are no longer identifiable for period cohort due to the presence of v_{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".

2.4.2 Resolving the curvature identifiability problem

Carstensen uses continuous functions to alleviate issues relating to the aggregation of years (Carstensen, 2007); however, he does not explore the curvature identifiability problem from the unequal intervals and whether continuous functions alone resolve this. Holford recognised the curvature identifiability problem (calling it the micro-trend identifiability problem) and described how it can produce an M-year cyclic pattern in the estimated temporal effects as well as proposing the use of smooth functions to model them (Holford, 2006). He argued that smooth functions resolve the curvature identifiability problem by providing sufficient structure to smooth over the cyclic nature. This latter proposal has not been fully explored and only demonstrated as a solution for a real-world data example.

When fitting APC models to unequal data, the curvature identifiability problem means the period and cohort curvature functions are no longer estimable Eq.(2.5). 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 $v_M \neq 0$, the period and cohort estimates will display the arbitrarily large, cyclic pattern over M-years as described by Holford (Holford, 2006). If $v_M \neq 0$ is approximated, the cyclic pattern in the period and cohort estimates is "wigglier" than when $v_M = 0$ is approximated. The PSS method we proposed in Section 2.2 has a penalty on the integrated square of the second derivative ("wiggliness") of the estimates. Therefore, an estimate for a function approximating $v_M \neq 0$ will have a larger integrated square of the second derivative and hence a larger penalty than one approximating $v_M = 0$. Consequently, the PSS method we propose will actively penalise the curvature identification issues; whereas, both Holford and Carstensen only smooth over them. A theoretical illustration of how the penalty function is alleviating the curvature identification problem can be found in Appendix A.

2.4.3 Simulation study

We now demonstrate this result empirically by repeating the simulation study from Section 2.3 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., $M = 5$).

The underlying single-year data are generated as described in Section 2.3. 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 $C' = M \times (A' - 1) + P$ 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 (Holford, 2006).

To get a true age effect that is comparable to the estimated age effect, we average the effect over every 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 $i = \{1, \ldots, A' = \frac{A}{M}\}$. 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 M cohorts (as age is aggregated in 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 $k = \{1, \ldots, C' = M \times (A' - 1) + P\}$. 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 2.3 are used. The models fit are the factor (FA), regression smoothing spline (RSS) and penalised smoothing spline (PSS) defined in Section 2.3. The estimated effects $\hat{h}^+_\star$ and curvatures $\hat{h}^+_{\star C}$ are calculated like Section 2.3 but with the vectors a' and c' for age and cohort; period is unchanged.

Figure 2-4 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 when the approximate function we are estimating is approximating $v_M = 0$, the estimates of the transformed period and cohort curvature functions are the smoothest. Therefore, the closeness between the RSS and PSS model estimates could be due to the cubic regression spline is approximating $v_M = 0$, which means the estimates are the smoothest and there is no additional penalty being incurred from the added identification.

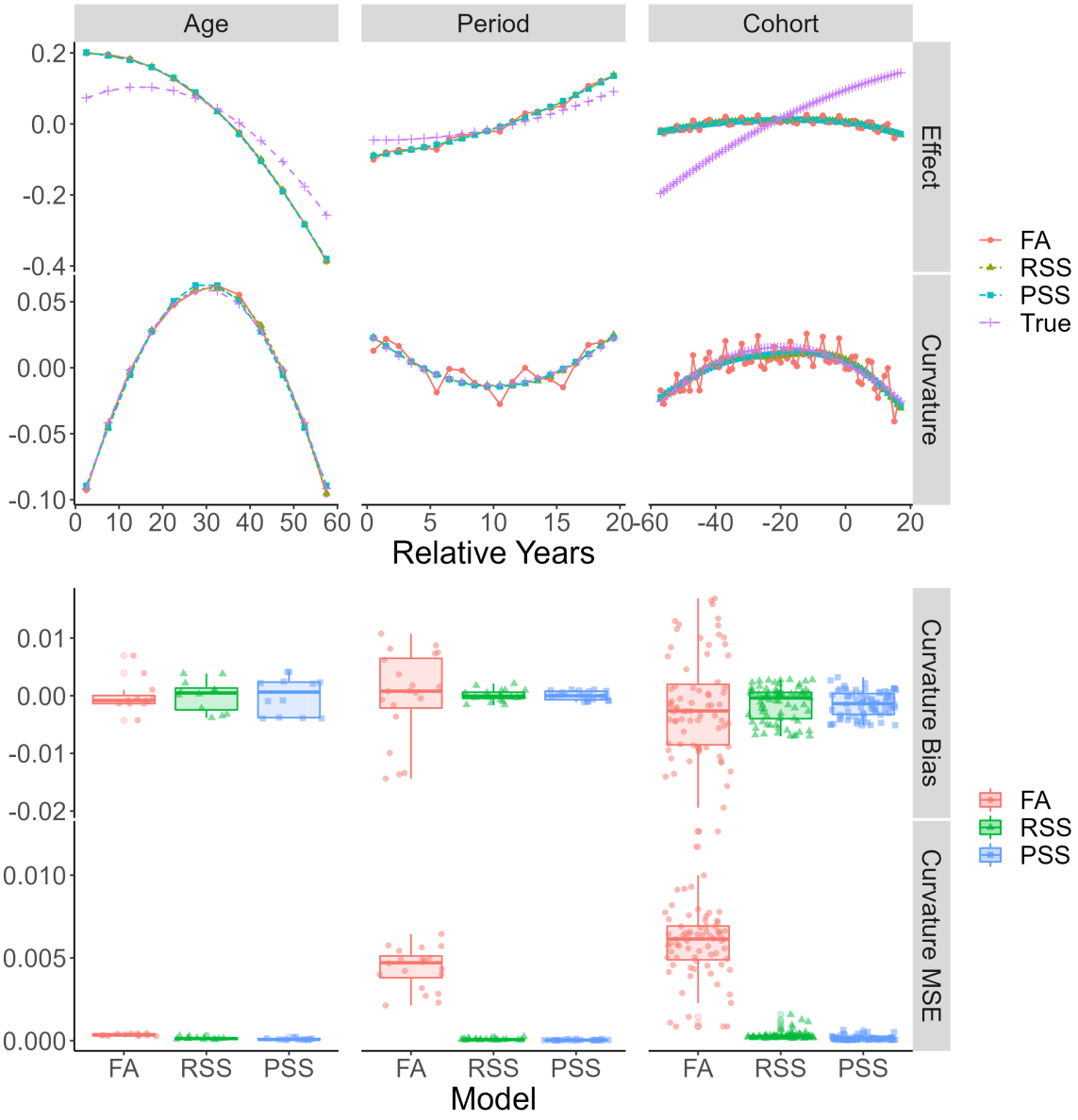


Figure 2-4: Simulation study results for unequal interval, $M = 5$, binomial data generated when all three temporal effects are present. The FA, RSS and PSS models are the factor, regression smoothing spline and penalised smoothing spline models, respectfully. The first and second row are of the temporal 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.

2.4.4 Sensitivity analysis

An important part of fitting APC models using splines is the basis and knot selection. Both Heuer (Heuer, 1997) and Holford (Holford, 2006) use one type of spline basis and give specific recommendations of knot specification. Heuer even caveats the recommended knot selection procedure with advice to test a range of choices for the number of knots to use. Due to the use

of only one basis function and the informed decisions needed to be made for the knot selection, we wish to test the robustness of the RSS and PSS models against the basis specification. 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 (Wood, 2017) 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 Figures 2-5 and 2-6, where the former is of the additional knots and the latter is of the 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.

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.

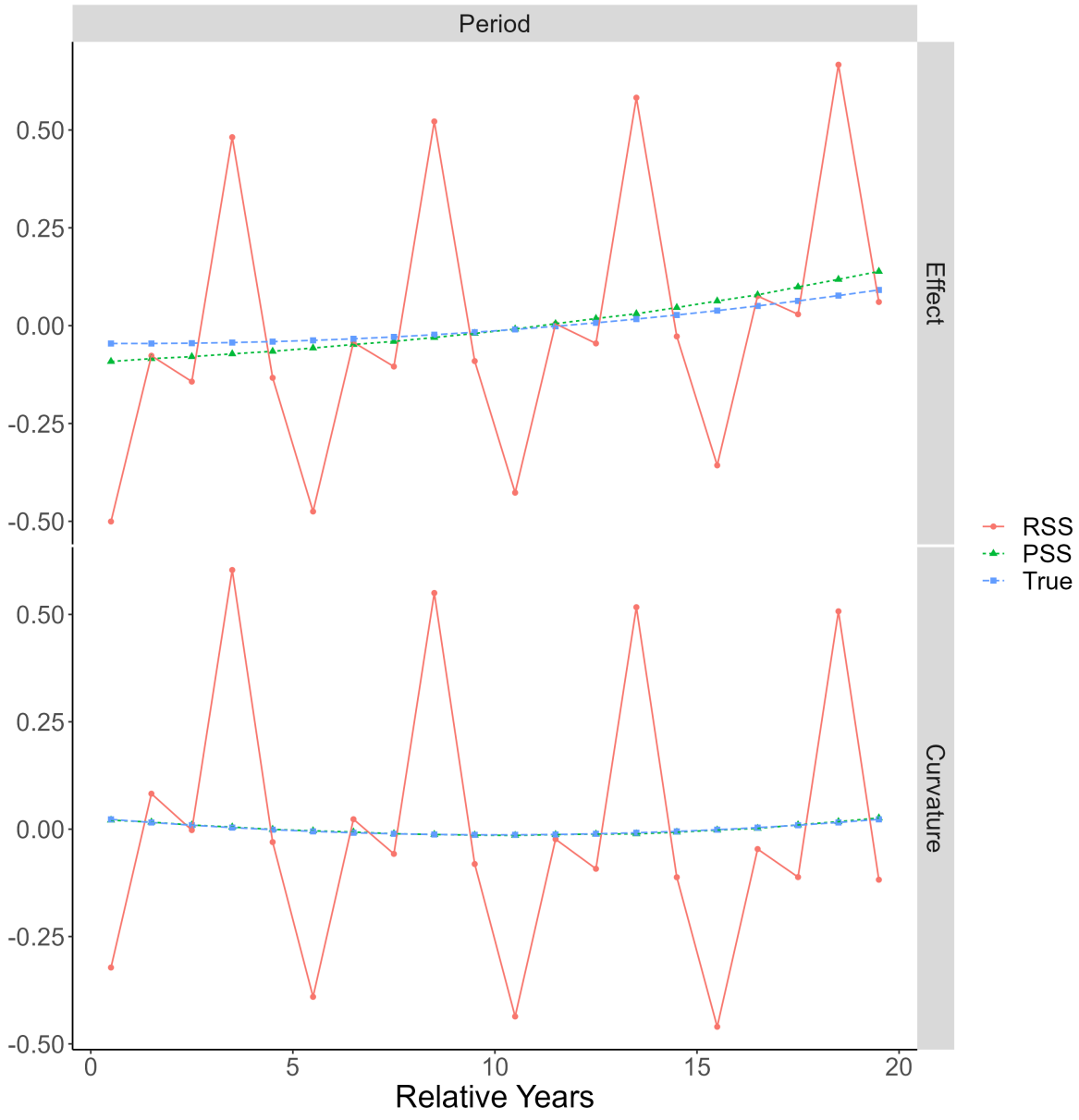


Figure 2-5: Simulation study results for the basis with additional knots for unequal interval, $M = 5$, binomial data generated when all three temporal effects are present.

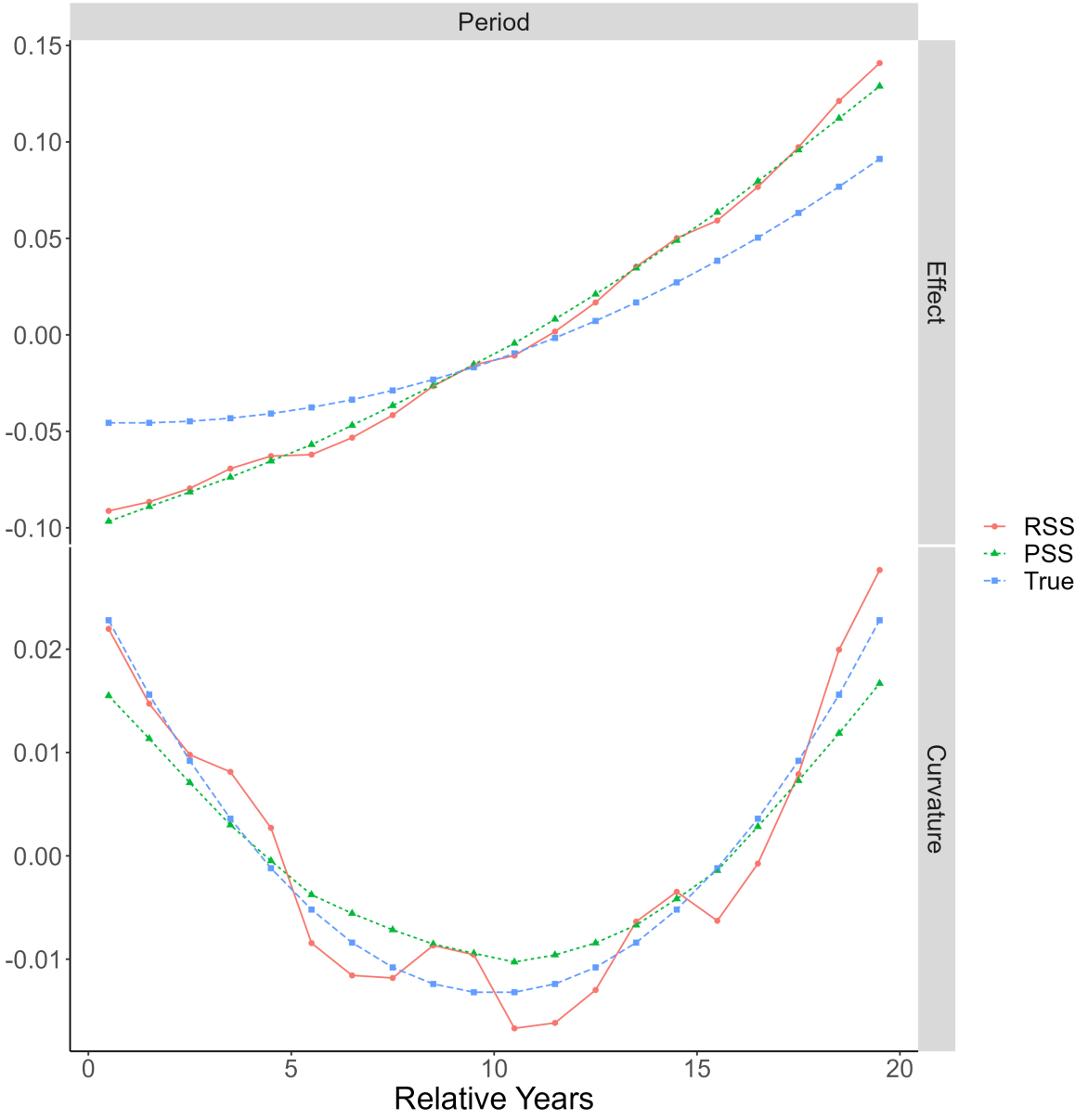


Figure 2-6: Simulation study results for the basis with additional periodic columns for unequal interval, $M = 5$, binomial data generated when all three temporal effects are present.

2.5 Application

We now consider an application of the PSS model on UK all-cause mortality data downloaded from the Human Mortality Database (HMD) (HMD, 2020). 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 (HMD, 2020)) which results in non-integer counts. For our models, we round the HMD values to the nearest integer.

Figure A-13 in Appendix A 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 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 2.3.

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

Table 2.3: UK all-cause mortality data aggregated in single-year age and period, five-year age and single-year period, and 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

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

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 A-14-A-16 in Appendix A 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 (Chien et al., 2015). 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 2-7 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 A-13 in Appendix A, 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 2-7. 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.

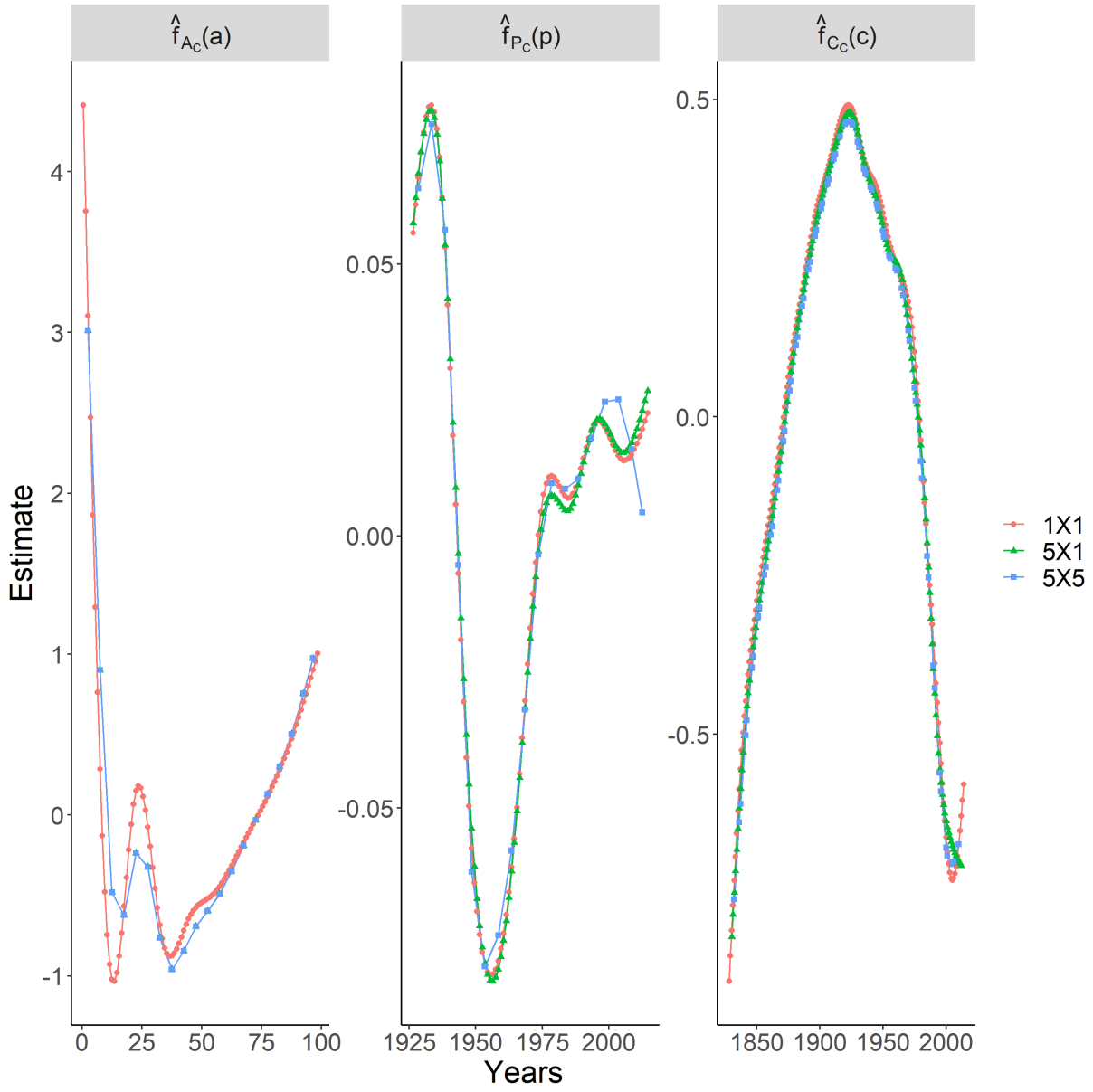


Figure 2-7: UK all-cause mortality fitted smooth curvatures for models fit to data aggregated in single-year age and period, five-year age and single-year period and five-year age and period.

2.6 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 (Holford, 1983) 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 a penalty is necessary to ensure alleviation of 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.

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 (Riebler and Held, 2010) or a Gaussian process prior (Chernyavskiy et al., 2018). These methods have parallels between our method due to the correspondence between penalised smoothing splines and stochastic processes (Wahba, 1978; Speckman and Sun, 2003) and we believe these provide suitable solutions. However, more work (see Chapter 3) 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 (Rosenberg, 2019). 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

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

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 (Kuang et al., 2008), 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 (ONS, 2020b); the first age group is split into $[0, 1)$ and $[1, 4)$. An additional example comes from the DHS (USAID, 2019), 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 (Etxeberria et al., 2017; Chernyavskiy et al., 2020) 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.

Supporting information

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

2.7 Chapter conclusions

The results of this chapter show the importance of including a penalty when fitting APC models to data that comes unequally aggregated. By using penalised smoothing splines, we described how to define and implement an APC model where the temporal terms and their respective penalties are reparameterised in the same manner. From a series of simulation studies and a sensitivity analysis, we compared the results of a factor model, an un-penalised smoothing spline model and a penalised smoothing spline model. The lack of robustness to both the way the spline was specified and the models ability at addressing the curvature identification problem in the un-penalised smoothing spline estimates in comparison to the penalised smoothing spline estimates confirmed the penalty is essential.

Whilst we used penalised smoothing splines to demonstrate the importance of a penalty term for the alleviation of the curvature identification problem, there are alternative methods to impose a penalty such as smoothing priors in a Bayesian paradigm. A correspondence between penalised smoothing splines and stochastic processes (such as those used as smoothing priors) has been noted in the spline literature (Wahba, 1978), but has not been considered within the APC literature. In Chapter 3, we consider this correspondence using random walk priors as the smoothing prior as these are commonly used when fitting APC models.

The application to all-cause mortality in the UK was used to highlight if our proposed model is appropriate for real-world data rather than give a detailed analysis of mortality in the UK. To provide a better example of how our model can be used in active research, in Chapter 4 we use our model to find estimates of under-five mortality rates (U5MR) with the goal of achieving the United Nations Sustainable Development Goal of reducing U5MR to 25 deaths per 1000 live births by the year 2030 (United Nations, 2019).

CHAPTER 3

COMPARING RANDOM WALK PRIORS TO PENALISED SMOOTHING SPLINES TO RESOLVE THE CURVATURE IDENTIFICATION PROBLEM

This chapter is a drafted manuscript with the purpose of understanding the relationship between a penalty imposed by either a penalised smoothing spline (PSS) or a smoothing prior. In particular, we focus on a random walk of order two (RW2) prior as they are commonly used within age-period-cohort (APC) models and have been used as a solution to the curvature identification problem.

A correspondence between penalised smoothing splines and stochastic processes (such as RW2 priors) can be explained using the more general theory of reproducing kernel Hilbert spaces. To understand this connection, one must understand and be comfortable with functional analysis. It is unlikely that practical users of PSS and RW2 prior models have this knowledge; therefore, this correspondence may have passed them. In a similar vein to Miller et al. (2020), who give an overview of the connection between PSS models and stochastic partial differential equations, we aim to give a similar exposition for PSS and RW2 prior models. As this manuscript is written for practical users, alongside the theoretical overview, we present a set of empirical results to highlight the correspondence. For APC modellers, we describe how to fit a reparameterised APC model with RW2 priors on the curvature terms and provide robust evidence that the correspondence holds when addressing the curvature identification problem for unequal interval data. We conclude by showing RW2 priors are imposing a penalty in an equivalent manner as the PSS model and can be used for APC models fit to unequal interval data.

Abstract

Models incorporating random walk of order two (RW2) priors have been used to smooth over cyclic patterns that arises in the temporal estimates of age-period-cohort models to data aggregated in unequal intervals. When doing so, it is common for the methods using RW2 priors to not acknowledge the curvature identification problem that causes the cyclic pattern. It has been shown previously that the penalty for deviations in linearity in the penalised smoothing spline (PSS) model is vital to fully address this problem. Both PSS and RW2 prior models are methods of imposing a penalty on deviations in linearity and there is a theoretical correspondence between them. In this paper, we include the key theoretical information that links these two approaches, and a set of empirical results so that a general practitioner can make the connection themselves and use the methods interchangeably.

3.1 Introduction

Non-parametric models are widely used since they offer a large amount of flexibility when compared to parametric models, by specifying the relationship between the observations and the covariates in terms of smooth functions. Here, we consider one function to be smoother than another if it is closer to linearity (less 'wiggly') than the other. Smoothness is defined with respect to linearity since we believe functions describing the relationship between dependent and independent variables are more likely to be linear than not. However, we do not want to enforce global linearity since this is extremely restrictive and rarely is the case. Therefore, we seek to find an estimate that is locally linear, i.e., it simultaneously promotes a smooth path between data points, whilst not being too aggressive and over-smoothing.

Two common ways of fitting smooth functions with a penalty to ensure the idea that straight lines are, generally, the smoothest fitting lines are: penalised smoothing splines (PSS) models and random walk (RW) prior models. There are theoretical parallels between estimates produced from PSS models and RW prior models based on theory from functional analysis. We wish to identify the key information from this theory so that a general practitioner can make the connection and use the methods interchangeably. In addition, the most appropriate software to implement each model is different and requires high-level technical knowledge to be able to customise.

In this chapter, we aim to address two key points: (i) understand the link between PSS model and the RW prior model, and (ii) show how to fit both models in their most appropriate software and see how closely related the results are. We fit both models in `R`, with the PSS model being fit using the `mgcv` package (Wood, 2017) and the RW prior model being fit using the `r-inla` package (Rue and Held, 2005).

Age-period-cohort (APC) models have been used to model incidence and mortality due to each of the temporal trends influence on a wide range of causes of death, but they suffer from a well-known structural link identification problem ($cohort = period - age$), and when fit to data that is aggregated into unequal intervals, the curvature identification problem also. In general, the most appropriate method to address the structural link identification problem is to reparameterise the model into identifiable quantities. For example, Holford (1983) reparameterises each temporal term into a linear trend and a set of curvatures that are orthogonal to both an intercept and their respective linear trend. Using a PSS model, we showed in Chapter 2 that a penalty is imperative to ensuring the curvature identification problem is alleviated. Alternatively, RW prior models have also been used as a solution for the curvature identification problem (Riebler and Held, 2010). Whilst understanding the relationship between PSS and RW prior models is important for all applied users of both models, those who fit APC models would benefit greatly from a more in depth understanding of the relationship since they are both solutions to the curvature

identification problem.

To address the first key point of this chapter, we give an overview of the correspondence between PSS models and RW prior models. Since we wish this to be for general applied users whom we assume do not have a background in functional analysis, we do not give a full, comprehensive breakdown, rather highlight the key correspondence, and then reference to literature for a more detailed consideration. The second key point we choose to address in a simulation study. A similar exposition on the relationship between penalised smoothing splines and models that use stochastic partial differential equations was conducted by Miller et al. (2020). For the APC model users, we further detail how to reparameterise both a univariate and full APC model which can be fit with RW priors and include simulation studies for both.

3.2 Penalised smoothing splines and random walks

For the purpose of explanation, consider a general univariate model,

$$y_i = f(x_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{3.1}$$

for observations y_i, a smooth function f of the covariate x_i that is unknown, and we wish to estimate, and $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ is the unstructured random (measurement) error. Currently, we consider our observations to be Gaussian but this can be extended to consider non-Gaussian observations using a link function, g, so the response is modelled with a specific distribution with the mean $g^{-1}(f(x))$. When estimating f, we can either define a finite dimensional approximation to f and then estimate the parameters of the approximation by maximising the penalised log-likelihood (as in the PSS model) or we can place a prior on f and define estimates of the model parameters by evaluating the posterior distribution via Bayes rule (as in the RW prior model).

By taking a partially informative Gaussian process (GP) prior on the function f, Kimeldorf and Wahba (1970), and later formalised by Wahba (1978), identified a correspondence between stochastic processes and a PSS. The penalised log-likelihood estimator of Eq.(3.1) is one that maximises

$$\hat{f} = \arg\max_f l(f) + \lambda \int f^{(p)}(x)^2 \, dx$$

where $l(f) = \log \mathcal{L}(f)$ is the log-likelihood and $\int f^{(p)}(x)^2 \, dx$ is a order-p penalty for the smooth f with the smoother parameter λ. For the cubic penalty seen throughout this thesis, $p = 2$. A partially informative GP prior on f is of the form,

$$\beta_0 + \beta_1 x + \cdots + \beta_{p-1} x^{p-1} + \delta^{\frac{1}{2}} Z(x)$$

where $\beta_0, \ldots, \beta_{p-1} \sim \mathrm{N}(0, \xi)$ as $\xi \to \infty$ which is "diffuse" (flat) on the coefficients that span

the null space of the penalty (for a cubic penalty, this is the constant and linear terms) and "proper" for the GP $Z(x)$ and δ is a scale parameter (Wahba, 1978). Wahba showed that the polynomial smoothing spline that maximises the penalised log-likelihood, $\hat{f}$, is equivalent to Bayesian estimation with a partially informative GP prior on f,

$$\hat{f} = \lim_{\xi \to \infty} \mathbb{E}_{\xi} \left[p(f) | \boldsymbol{y} \right]$$

where $\xi \to \infty$ corresponds to the "diffuse" part of the prior. Therefore, the correspondence is between the polynomial smoothing spline maximiser of the penalised log-likelihood and the posterior expectation of the Gaussian random field defined by a partially informative GP prior. This result was later extended by Speckman and Sun (2003), who showed the class of priors that are appropriate for the correspondence includes those motivated by a difference approximation in the penalty for the PSS. For example, by taking a discretised version of the penalty (second differences), the prior leads to a discretised Bayesian smoothing spline estimate where the prior is a RW of order two (RW2) prior.

To see the relation between a PSS and a RW2 more clearly, recall from Chapter 1 we can represent f in Eq.(3.1) by a smooth spline such that for the basis $f(x) = \sum_{k=1}^{K} b_k(x) \beta_k$, the penalty function can be expressed as $\int f''(x)^2 dx = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$, where $\boldsymbol{S}$ is known as the penalty matrix. If we discretised the penalty and take second differences, the penalty can be expressed,

$$\int f''(a)^2 da = \sum_{k=1}^{K-2} \left(f(x_k) - 2f(x_{k+1}) + f(x_{k+2}) \right)^2 = \sum_{k=1}^{K-2} \left(\beta_k - 2\beta_{k+1} + \beta_{k+2} \right)^2 = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \tag{3.2}$$

where the discretisation of the penalty and finite approximation were used after the first and second equalities, respectfully. Given the finite approximation of the true function, an estimate can be found by maximising the following penalised log-likelihood,

$$l_p(\boldsymbol{\beta}, \boldsymbol{\theta}) = l(\boldsymbol{\beta}, \boldsymbol{\theta}) - \lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$$

where $l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta})$ for the likelihood function $\mathcal{L}$ and the finite basis approximation to the true function is used. To see the relationship between a PSS and RW prior model, rewrite the penalised log-likelihood in terms of a likelihood,

$$\mathcal{L}_p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) \times \exp\left(-\lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \right).$$

If we consider the key concept in Bayesian inference (*posterior* $\propto$ *prior* $\times$ *likelihood*), $\mathcal{L}_p(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta})$ are equivalent to the posterior distribution and the likelihood function, respectively and $\exp\left(-\lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \right)$ can be thought of as the prior distribution of the parameters $\boldsymbol{\beta}$. Given the

second difference penalty, Eq.(3.2), the prior distribution

$$p(\boldsymbol{\beta}|\lambda) \propto \exp\left(-\boldsymbol{\beta}^T \boldsymbol{Q}_\lambda \boldsymbol{\beta}\right)$$

where $\boldsymbol{Q}_\lambda = \lambda \boldsymbol{S}$ and this is of the form of an improper Gaussian Markov random field (GMRF) prior on a regular lattice (Wood, 2017; Yue et al., 2012). The impropriety in the prior is due to the penalty matrix $\boldsymbol{S}$ not being of full rank, meaning the precision matrix $\boldsymbol{Q}_\lambda$ is not invertible. The precision matrix is rank deficient by the dimension of the null space for the penalty matrix $\boldsymbol{S}$.

As described in Chapter 1, a RW of order p prior is an improper Gaussian prior (with a Markov property) which has a rank deficiency of p. If we choose to use a RW2 prior, then $\boldsymbol{Q}_\lambda$ will be an improper GMRF with a rank deficiency of two. This is equivalent to using the penalty function since $\boldsymbol{S}$ is rank deficient by two. Because of this, both the RW2 prior and the penalty are penalising deviations in linearity.

Whilst both the PSS model and the RW2 prior model are imposing a penalty on the second derivative of the estimates, how each model performs this differs which causes slight differences in practical estimates. For example, the smoothing parameter of the PSS model is estimated by cross validation in `mgcv` (Wood, 2017) whereas the equivalent smoothing parameter of the RW2 prior model in `r-inla` is based upon the choice of the prior distribution (Rue and Held, 2005). The data driven approach of the PSS model is attempting to find an 'optimal' smoothing parameter; whereas in the RW2 prior model, we specify a distribution for the smoothing parameter *a priori*.

Having identified the key information required to understand that a PSS model and RW2 prior model can be used interchangeably, we want to see if this is still true when fitting APC models to data unequally aggregated. To do this, we first define a reparameterised APC model. The reparameterisation for a PSS model can be found in Chapter 2. Therefore, in the next section, we focus on how the reparameterisation works for a RW2 prior model.

3.3 Random walk priors for age-period-cohort modelling

APC models suffer from the structural link identification problem since $cohort = period - age$. Because of this linear relationship between the temporal terms, when including all three together, the model becomes overparameterised and each of the terms are unidentifiable. Therefore, we can add and subtract a linear trend to each of the temporal terms without changing the overall linear predictor. The structural link identification problem is often resolved by reparameterising the model in terms of identifiable quantities. We choose to reparameterise each temporal term into a linear trend and a set of curvatures that are orthogonal to the linear trend (Holford, 1983).

In a Bayesian paradigm, having a fully identifiable model is of less concern if the linear predictor (e.g., mean or log rate) is identifiable since the posterior distribution can still be sampled from (Besag et al., 1995). However, it is best practise to always ensure all terms in the model are identifiable, not just the linear predictor (Smith and Wakefield, 2016).

When an APC model is fit to data that comes in unequal intervals, the previously identifiable temporal curvatures (or an equivalent) become unidentifiable, hence the term 'curvature identification problem'. Due to this, there appears a cyclic pattern in the period and cohort estimates (both for the full effect and the curvature terms) (Holford, 2006; Held and Riebler, 2012). We showed in Chapter 2 that a penalty on the curvature terms was a necessary inclusion to resolve the problem. In a similar vein, Riebler and Held (2010) imposed a RW2 prior on each of the temporal terms to smooth out the cyclic pattern, thereby addressing the problem. Even though the cyclic pattern is resolved, this is a symptom of the problem and the actual problem, the now unidentifiable curvature function, is not acknowledged in this work.

A model that imposes a RW2 prior penalises deviations from linearity to promote a smooth fitting estimate in a similar way to the PSS model. Having described the correspondence between PSS models and RW prior models, we look to see if this correspondence is affected by the issues relating to APC models specifically. Namely, the structural link and curvature identification problems. We do this empirically with simulated data by fitting both models to the same data and then comparing the estimates against the truth, to see how well the models are addressing the identification problems, and against one another, to see how closely related the two models are. Before the simulation study, we first describe how to include the RW priors in a reparameterised APC model.

3.3.1 Orthogonalization with random walk priors

Consider again the simple univariate model Eq.(3.1) where f has a prior distribution which is a RW2 prior, i.e., $p(f)|\tau \sim \text{RW2}(\tau)$. The precision matrix of an RW2 prior is rank deficient meaning the RW2 prior does not have a proper multivariate normal distribution. The impropriety of the GMRF is addressed by conditioning the improper distribution on a set of constraints, $\boldsymbol{Cx} = \boldsymbol{0}$, where $\boldsymbol{C}$ is a constraint matrix defined by the vectors that span the null space of the precision matrix. Consequently, the improper GMRF is equivalent to a proper GMRF on a lower dimension when conditioned by some linear constraint. To give more detail on this, we now describe the constraints that are implicit on improper GMRFs. Our description follows the general case of a RWp prior that can be found in Rue and Held (2005), but we consider a RW2 prior.

To be explicit in the following, let us redefine the RW2 prior as $p^*(f)|\tau \sim \text{RW2}(\tau)$ which has a precision matrix $\boldsymbol{Q}^*$. The $*$ is used to explicitly denote the improper density and rank

deficient precision matrix for this explanation only. The precision matrix can be eigendecomposed into $\boldsymbol{Q}^* = \boldsymbol{V}\boldsymbol{\Lambda}^*\boldsymbol{V}$ where $\boldsymbol{\Lambda}^* = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_n)$ is a diagonal matrix of eigenvalues and $\boldsymbol{V}^* = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_n)$ is the corresponding eigenvectors. Since $\boldsymbol{Q}^*$ is of rank $n-2$, there are two zero eigenvalues, say, λ_1 and λ_2, with $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and their respective eigenvectors. These eigenvectors are said to span the column space of $\boldsymbol{C}^T$ (the null space of $\boldsymbol{Q}^*$) and correspond to the constant and linear functions. Therefore, the density of the RW2 prior on f is,

$$p^*(f) = p(f|\boldsymbol{C}\boldsymbol{x} = \boldsymbol{0}) \text{ where } \boldsymbol{C} = \{\boldsymbol{1}, 1:n\}$$

and the density of the improper GMRF is equivalent to the density of a proper GMRF which has been conditioned to be invariant to the addition of the functions that span the null space of the improper GMRF. Consequently, imposing a RW2 prior on a term is equivalent to removing a constant and linear trend from it, which enforces both sum-to-zero and zero-slope constraints. In addition to the constraints implied on the term, the null space of precision matrix for the prior is spanned by a constant and linear term which is the same as the penalty matrix for the penalised smoothing spline. Therefore, the RW2 prior is penalising deviations in linearity like the penalised smoothing spline.

In terms of the age only model, let $a = 1, \ldots A$ and consider the function of age

$$\eta_a = f_A(a)$$

where η_a is the linear predictor and $f_A|\tau_a \sim \text{RW2}(\tau_a)$ is an age term with a RW2 prior. We define a reparameterised age model with terms for an intercept, linear trend and curvature,

$$\eta_a = \beta_0 + a\beta_1 + f_{A_C}(a)$$

where β_0 and β_1 are the parameters for an intercept and slope, respectively, and $f_{A_C}|\tau_a \sim \text{RW2}(\tau_a)$ is the curvature term with a RW2 prior. Identifiability of the model is ensured through sum-to-zero constraints on both the linear term and the curvatures and additional zero-slope constraints on the curvature. The sum-to-zero and zero-slope constraints on f_{A_C} are, as we saw, enforced when we impose a RW2 prior which means the f_{A_C} term is invariant to the constant and linear terms and is penalising deviations from linearity in the model estimates.

Based on the reparameterised age model, a full APC model can reparameterised similarly. Let $a = 1, \ldots, A$ and $p = 1, \ldots P$, a reparameterised APC model is

$$\eta_{ap} = \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c)$$

where t_1 and t_2 are two of three temporal trends, β_1 and β_2 are arbitrary parameters and $f_{A_C}|\tau_a \sim \text{RW2}(\tau_a)$, $f_{P_C}|\tau_p \sim \text{RW2}(\tau_p)$ and $f_{C_C}|\tau_c \sim \text{RW2}(\tau_c)$ are the curvature terms each with a RW2 prior.

For t_a, t_p and t_c, the linear trend for each of the temporal terms, any combination of $\kappa_1 t_a + \kappa_2 t_p + (\kappa_2 - \kappa_1)\, t_c$ is estimable (Holford, 1983). When we say 'dropping the linear trend for' (or equivalent) we are referring to the choice of κ's. For example, if we chose $\kappa_1 = \kappa_2 = 1$, we retain the linear trends for age and period and drop cohort. The cohort trend is not completely omitted from the model, rather subsumed by the age and period trends, i.e., the parameter for age will reflect both age and cohort and the parameter for period will reflect both period and cohort.

3.4 Simulation Study

We now conduct three simulation studies. The first study is to show an empirical understanding of the correspondence between the PSS model and the RW2 prior model. We fit the same simple univariate (SU) model implemented either by a PSS model or a RW2 prior model. The second and third studies are more aimed at the APC modellers. The second study is for a reparameterised univariate (RU) model and shows if the reparameterisation causes any adverse effects on the correspondence and model fitting process. The final study uses a full APC model to see if a RW2 prior model is appropriate for addressing the curvature identification problem.

The simulated data is recreated from the unequal intervals data from the Chapter 2 simulation. For a quick recap, the original temporal functions are from Luo and Hodges (2016) but were adapted to reflect the UKs National Health Services obesity survey (NHS, 2020). We simulate data for single-year ages between $[0, 60]$ and single-year periods between $[0, 20]$. We then collapse age over five years to form the unequal intervals and define cohort using $cohort = period - age$. The data was generated like this to replicate how aggregated data is normally collected in single years and then collapsed over. For modelling, we take the mid-point of each of the groups and model them as continuous values, i.e., for the ages, we take $2.5, 7.5, \ldots, 57.5$. In Chapter 2 we simulated from three different data sets, binomial, Gaussian, and Poisson. The qualitative results from each distribution are the same; consequently, we choose to perform the simulations in this chapter for binomial data only. The data for the SU and RU models is generated using the age function only, and the data for the APC model is generated using all three temporal functions.

Let $a = 1, \ldots, A$ and $p = 1, \ldots, P$ be the age and period indices. The three models are defined

$$\begin{aligned}
\text{SU}: \quad \eta_a &= \beta_0 + f_A(a) \\
\text{RU}: \quad \eta_a &= \beta_0 + a\beta_1 + f_{A_C}(a) \\
\text{APC}: \quad \eta_{ap} &= \beta_0 + t_1\beta_1 + t_2\beta_2 + f_{A_C}(a) + f_{P_C}(p) + f_{C_C}(c)
\end{aligned}$$

where η is the linear predictor, β_0 is the intercept, t_1 and t_2 are two of three temporal trends, β_1 and β_2 are arbitrary parameters, f_A is the full age term and f_{A_C}, f_{P_C} and f_{C_C} are the temporal curvature terms. The full age and temporal curvature terms are implemented as either

a smoothing spline (in the PSS model), or with a RW2 prior (in the RW2 prior model).

3.4.1 Implementation

The RW2 prior models will be fit using integrated nested Laplace approximations (INLA) as implemented in the `r-inla` package Rue et al. (2009). INLA provides accurate approximations of the marginal posterior distribution for random effects and hyper-parameters whilst avoiding the need for costly and time-consuming Markov-chain Monte Carlo (MCMC) sampling that is accustomed to Bayesian models. The PSS models will be fit using the `mgcv` package (Wood, 2017). To see how to implement both models, consider an example for a model with one parametric component (`x1`) and one non-parametric component (`x2`)

In `r-inla`, for a non-parametric component with a RW2 prior, the model formula is $\texttt{y} \sim \texttt{x1} +$ `f(x2, model = 'rw2', ...)` where `f()` is to call a smooth function. Within `f()`, the only argument that needs to be explicitly defined to fit a RW2 prior model is `model = 'rw2'`. Other additional arguments that are commonly used are: `extraconstr = list(A, e)` where $\texttt{A} = \boldsymbol{C}$ is the constraint matrix and $\texttt{e} = \mathbf{0}$ to specify the constraint explicitly, and `hyper` for the specification of the prior specification. Part of a Bayesian model fitting process is the prior specification which we speak about in Section 3.4.2.

In `mgcv`, for a non-parametric component approximated with given basis (we choose to use a cubic regression spline basis), the model formula is $\texttt{y} \sim \texttt{x1} +$ `s(x2, bs = 'cs', k = x2k)` where `s()` is to call a smooth function. Within `s()`, the `bs` argument is for the choice of spline basis used, we use a cubic regression spline but `mgcv` has several bases that can be implemented, see Wood (2017), and the argument `k` is used to specify the number of knots used to define the spline basis. For the orthogonalized PSS model, the orthogonal basis and penalty can be defined explicitly by making use of `mgcv`'s `fit = FALSE` argument in the `gam()` call. This returns the model and penalty matrices before the model fitting process, allowing us to make any explicit changes prior to the model fitting. We use the method previously described in Chapter 2 to orthogonalize these matrices and perform the model fitting process.

3.4.2 Prior specification and sensitivity

Within the model prior choice there are parameter(s), and as a part of a Bayesian hierarchical model, we define a prior distribution on said parameter(s) which can be called a hyperprior. For example, the RW2 prior model has the precision parameter τ and in `r-inla` we must specify a hyperprior distribution on the precision. In addition, the hyperprior distribution will depend on a set of parameters call hyperparameters that we must specify. In Bayesian analysis, the results attained may depend on the underlying hyperprior specification. To account for this, it is

important to perform a sensitivity analysis on whether the results change substantially depending on prior specification. We perform the sensitivity analysis using the SU model. Alongside the default `r-inla` hyperprior and hyperparameter specification for a RW2 prior model, we test two different distributions each with three sets of hyperparameters. We do not wish to find the best possible combination of hyperprior and hyperparameters, rather show if there is a substantial difference between different specifications. For that reason, we make our choices below based off order of magnitude changes.

One possible choice of hyperprior distribution is the penalised complexity (PC) priors (Simpson et al., 2017). PC priors are designed to penalise deviations in the posterior distribution from a simpler, "base" model. PC priors were developed as a solution to the problem that arises when a prior forces overfitting (when it is impossible to distinguish between a model that is supported by the data or by a poor prior choice). For ζ, a flexibility parameter controlling how much the model results depend on the data and the prior, $\mathcal{Q}(\zeta)$ is an interpretable transformation (i.e., standard deviation, precision) of ζ. In addition, let U be a "sensible" upper bound and p the probability of ζ being in the upper bound, then a PC prior is specified

$$\mathrm{P}(\mathcal{Q}(\zeta) > U) = p.$$

The default hyperprior for the precision parameter in `r-inla` for a RW2 prior model is a Gamma prior specified with a scale and rate parameter, $\mathrm{Ga}(scale, rate) = \mathrm{Ga}\left(1, 5 \times 10^{-5}\right)$. To keep with the tradition of using Gamma priors for a RW2 prior precision parameter but wishing to use PC priors to penalise an incorrect prior specification, we first specify a set of three PC priors and then define a Gamma prior that has an equal mean "mean matched" to each of the three PC priors alongside the default `r-inla` choice.

To make the Gamma prior "mean matched" to the PC prior:

1. Take k samples from the PC prior distribution

$$z_i \sim \mathrm{P}(\tau > U) = p$$

2. Find the empirical mean of the PC prior distribution:

$$\frac{1}{k}\sum_{i=1}^{k} z_i$$

3. Using a fixed scale parameter and the theoretical expectation of the Gamma distribution, $scale/rate$, calculate the rate parameter to ensure the Gammas distribution expectation

matches the empirical mean of the PC priors distribution

$$rate = scale/\left[\frac{1}{k}\sum_{i=1}^{k} z_i\right]$$

We fix the probability in the PC prior $p = 0.01$ and change the upper bound by orders of 0.5, $u = 1, 1.5, 2$. Fixing the scale parameter, $scale = 1$, we calculate the Gamma distributions rate parameter for each choice of PC prior using the above process. The full set of hyperprior distributions are shown in Table 3.1.

Name	Prior Choices
Default	$\text{Ga}\left(1, 5 \times 10^{-5}\right)$
PC1	$\text{PC}\left(\tau > 1\right) = 0.01$
PC2	$\text{PC}\left(\tau > 1.5\right) = 0.01$
PC3	$\text{PC}\left(\tau > 2\right) = 0.01$
Ga1 (Match PC1)	$\text{Ga}\left(1, 6.58 \times 10^{-6}\right)$
Ga2 (Match PC2)	$\text{Ga}\left(1, 4.43 \times 10^{-7}\right)$
Ga3 (Match PC3)	$\text{Ga}\left(1, 1.01 \times 10^{-6}\right)$

Table 3.1: Prior specifications for the random walk 2 model used in the sensitivity analysis.

The results of the SU simulations for these specifications are in Figure 3-1. The bias and mean square error (MSE) boxplots are all relatively similar, with the width of the inter-quartile range (IQR) being small and mean being close to zero. Whilst the y-axis scale of the boxplots indicate how marginal the differences are, the PC priors have a smaller MSE and a narrower bias IQR than the Gamma priors. Therefore, we use the PC priors for the following simulation studies. In particular, we use the PC1 prior, but there is little to distinguish between the PC1, PC2 and PC3 prior results.

3.4.3 Results

3.4.3.1 Simple univariate model

The results from the SU simulations comparing the PSS model to the RW2 prior model with a PC1 prior are summarised in Figure 3-2. The first row shows the true and estimated temporal effects and the second shows the true and estimated temporal curvatures. The bottom two rows are the bias and MSE boxplots for the curvature estimates. In APC modelling, the true effects are unidentifiable, but the curvatures are. Therefore, to appropriately assess a full APC model, we want to evaluate the plots based on identifiable terms. Hence why the bias and MSE are based off the curvature. As there are no structural link in the age only model (as only one temporal effect is being considered), both the full effects and curvatures are identifiable. This contrasts

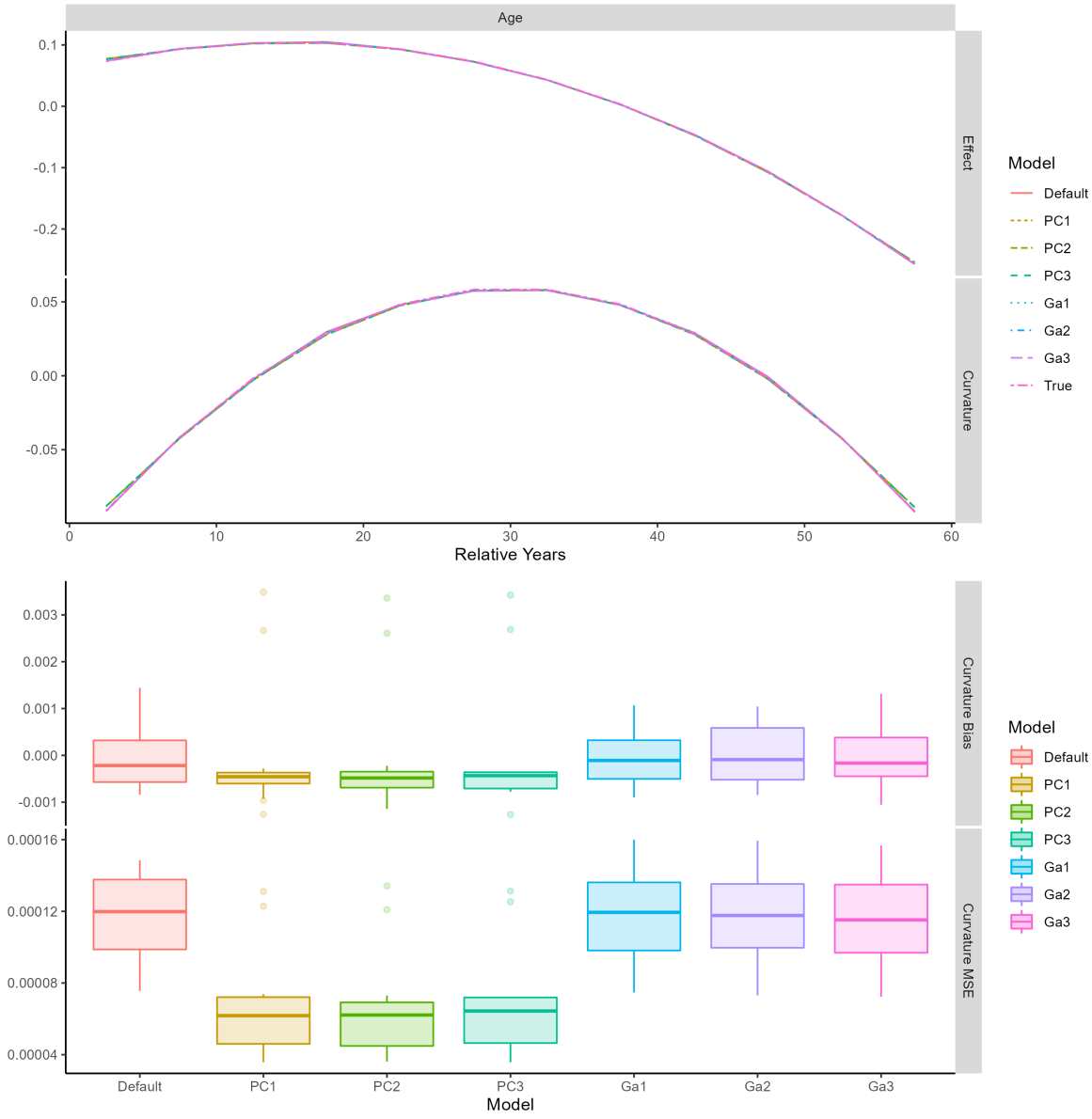


Figure 3-1: Simulation study results for the sensitivity analysis on the simple age model fit with RW2 priors onto binomial data. The first and second rows are of the temporal effect and curvature for both the models alongside the true values. The bottom two rows are the bias and MSE box plots of the curvature for each model.

with an APC model which has the structural link so only the curvatures are identifiable.

Due to the lack of structural link in the age only model, both the full effects and the curvatures are identifiable, and this is shown in Figure 3-2 by the PSS model and RW2 prior model estimates both matching the truth. The IQR of the bias from the RW2 prior is much smaller than that from the PSS model as well as being marginally closer to zero. For the MSE, the IQR of both models is similar with the PSS model having an average MSE marginally closer to zero than the

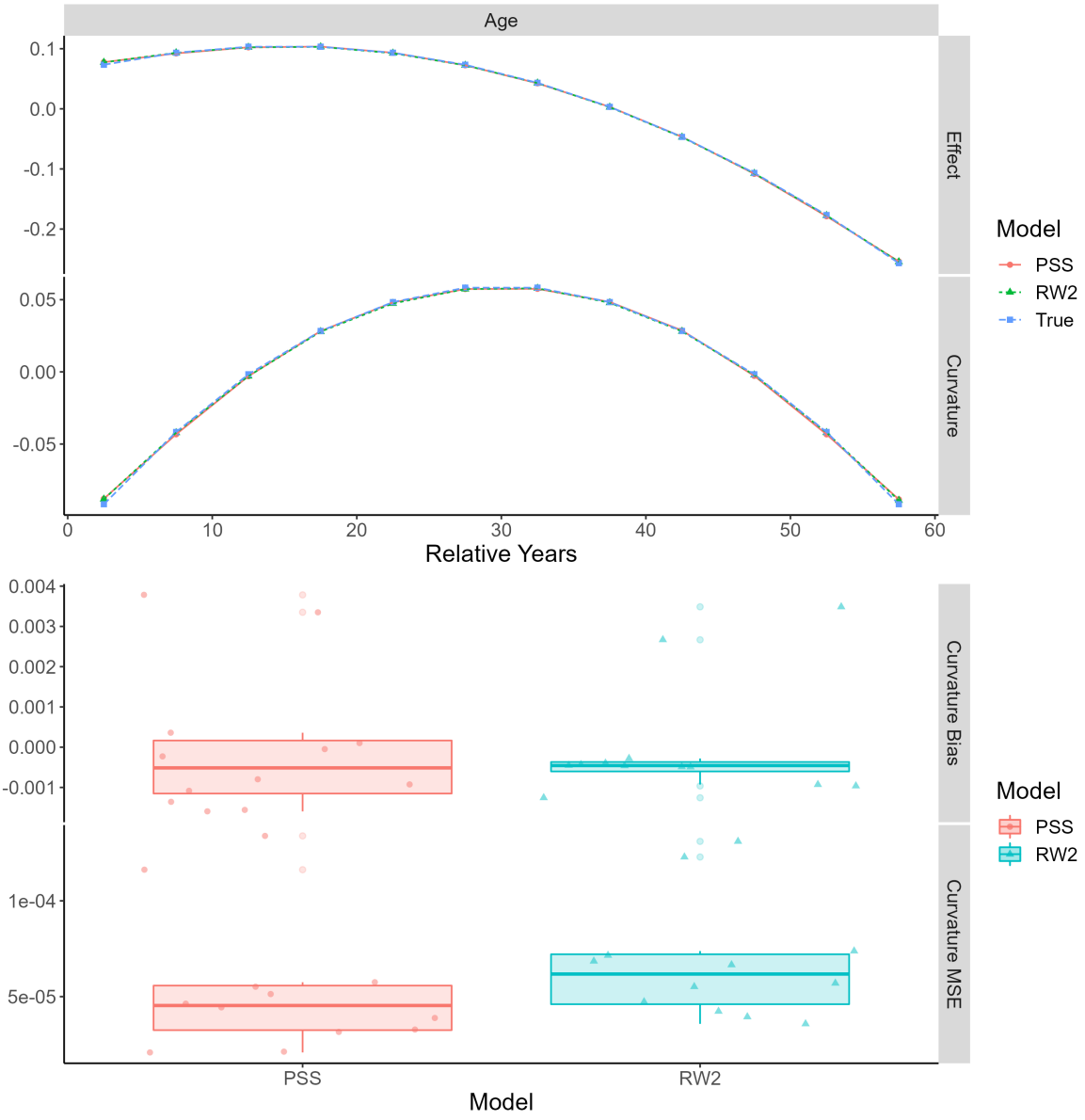


Figure 3-2: Simulation study results for the simple age model fit with a PSS and RW2 prior model onto binomial data. The first and second rows are of the temporal effect and curvature for both the models alongside the true values. The bottom two rows are the bias and MSE box plots of the curvature for each model.

RW2 prior model, see the scale. If we chose to use a Gamma prior, we might see a slightly larger difference in the MSE between the PSS and RW2 prior models as indicated by the sensitivity analysis, Figure 3-1.

An important goal of the simulation study is to check whether a RW2 prior model can be used in the place of a PSS model when smooth estimates are needed. To better see how close the results of the PSS and RW2 prior models are, we compare them directly in Figure B-1 in Appendix B.

In Figure B-1, all the points fall along the $y = x$ axis, showing that both sets of estimates are similar. As expected from the theoretical results that highlighted the correspondence between the PSS model and the RW2 prior model, the results from the simulation study on the SU model confirmed that the models are interchangeable with one another as well as both being appropriate for fitting a smooth curve through the data.

3.4.3.2 Reparameterised univariate model

When fitting the full APC model, we reparameterise each of the temporal terms into a linear trend and a set of orthogonal curvature terms to resolve the structural link identification problem. To ensure this reparameterisation does not adversely affect the correspondence between the PSS model and the RW2 prior model, and not have the complication of the structural link that would be present in the APC model, we include the RU model simulation study. The results of the RU study are in Figure 3-3 for the model comparison against the truth and Figure B-2 in Appendix C for the comparison between models. As expected, the reparameterisation had no major change on the overall outcome that both models can be used interchangeably and are effective for fitting a smooth function.

In the RU model results, there is a slight difference between the PSS estimates and the truth that was not present in the PSS SU model results. This is highlighted by the PSS RU models bias IQR being larger than the PSS SU models bias IQR. Since the smooth function basis only contains those bases relating to the curvature, it is only capturing the higher frequencies so will incur a larger penalty. When penalising, the PSS model is balancing a trade-off between an increased bias and smaller variance, hence the larger bias IQR.

3.4.3.3 Age-period-cohort model

Figure 3-4 shows the results of the APC model. The rows are the same as in the previous figures, but we now have three columns, one for each of the temporal effects, age, period, and cohort (left-to-right). The first noticeable difference between the APC model results and the SU and RU model results is that the age full effect for each of the APC models are no longer the same as the true effect or one another. Unlike in the SU and RU models where there is only one temporal term, the APC model has all three terms and hence will suffer from the structural link identification problem, making the full effects unidentifiable. The differences between all three sets of effects are due to this. The lack of identification in the full effects is the reason we include estimates of the curvatures as well as basing any validation from them.

Considering the curvature estimates, bias and MSE. Neither set of estimates shows signs of the cyclic pattern from the curvature identification problem, indicating they are both addressing the

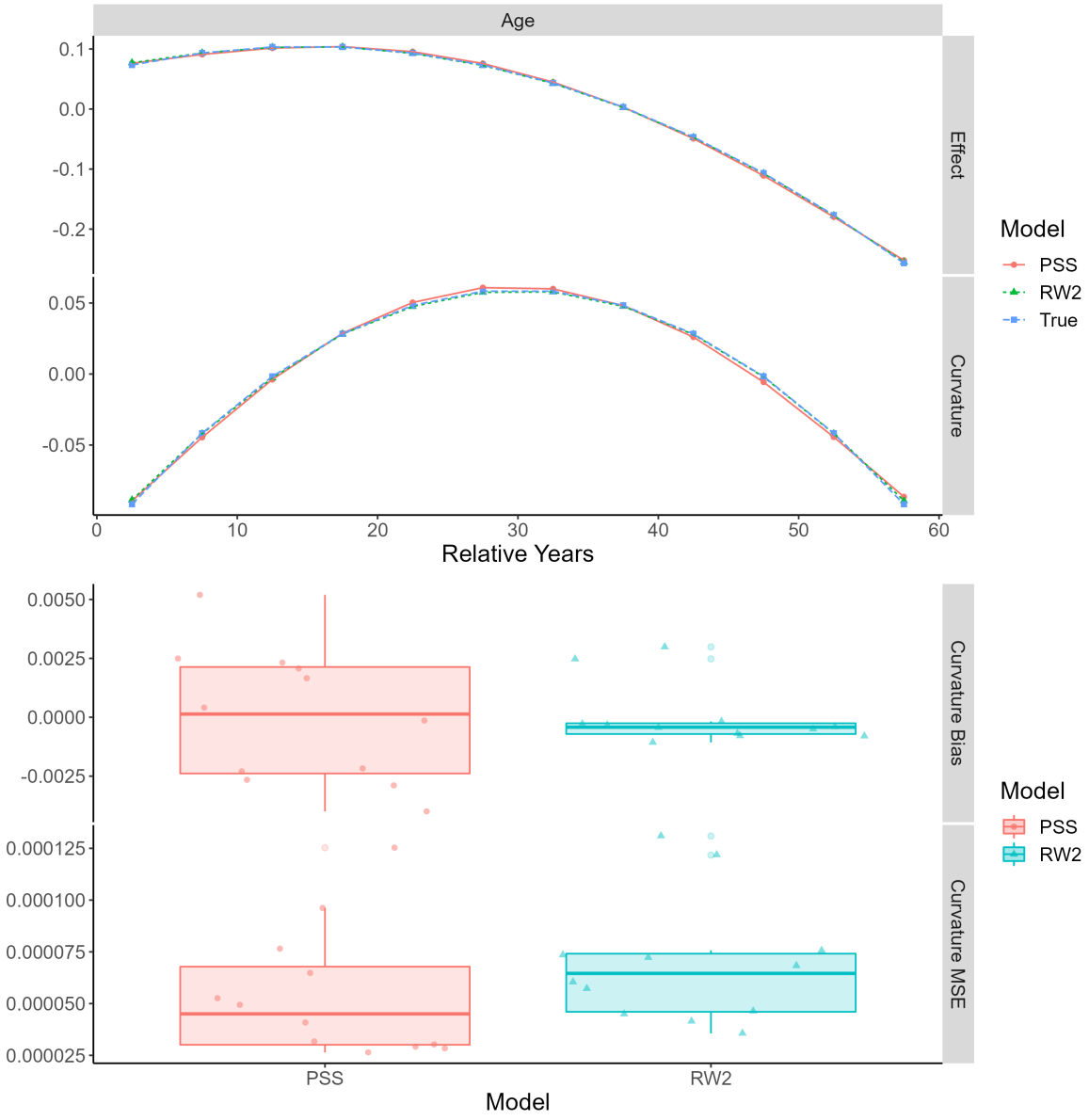


Figure 3-3: Simulation study results for the orthogonalized age model fit with a PSS and RW2 prior model onto binomial data. The first and second rows are of the temporal effect and curvature for both the models alongside the true values. The bottom two rows are the bias and MSE box plots of the curvature for each model.

problem as expected. We know the suitability of the PSS model is due to the penalty and we know there is the correspondence between the PSS and RW2 prior models; therefore, we conclude the suitability of the RW2 prior model is due to it enforcing its own penalty to deviations in linearity in the estimates. Both the period and cohort curvature estimates for the RW2 prior model appear to be slightly flatter than the truth and the PSS model estimates. This can be attributed to two things: firstly, in general, Bayesian models tend to attenuate estimates (Smith and Wakefield, 2016) and more extreme values attenuated to a greater extent results in a flatter

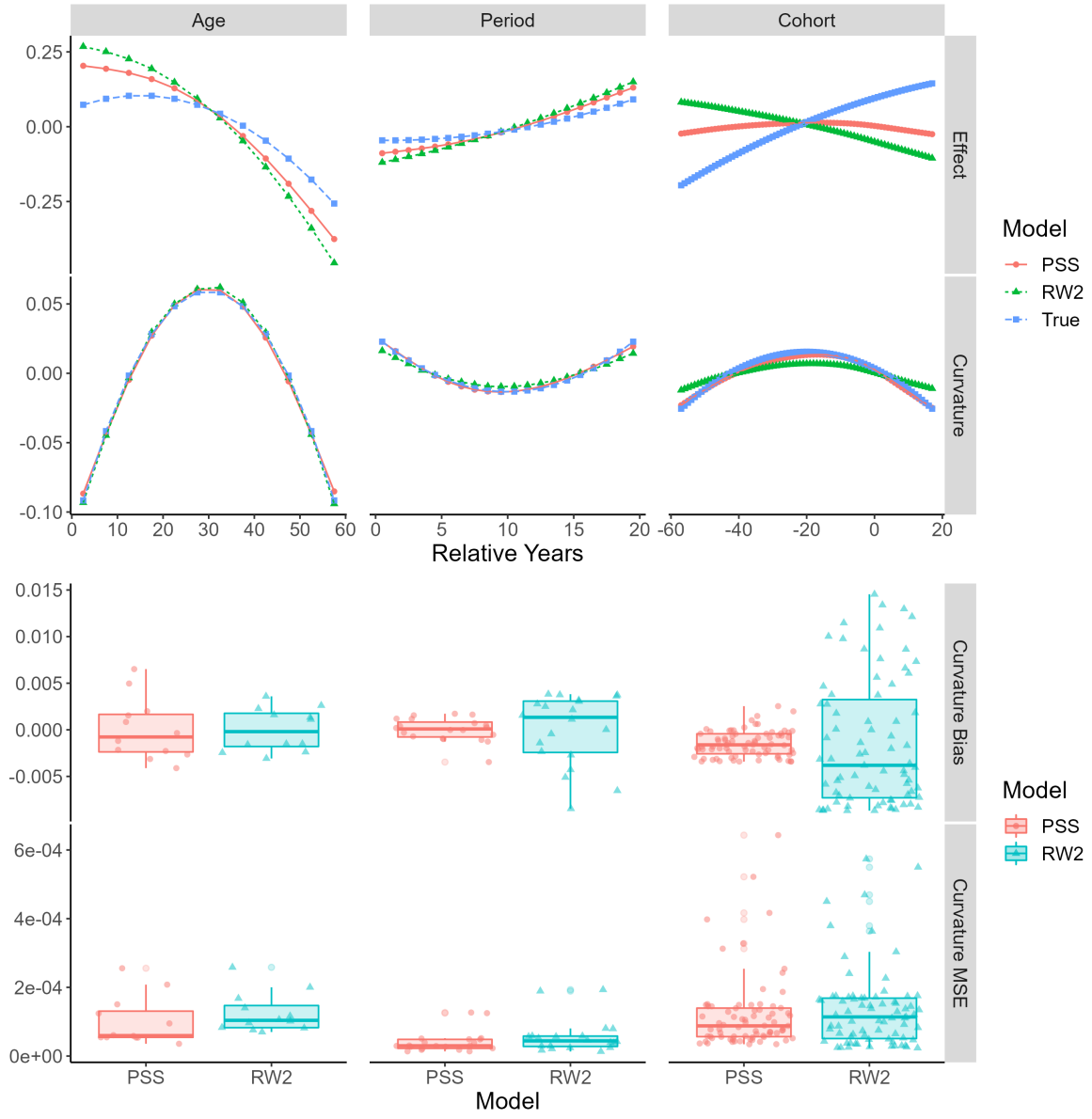


Figure 3-4: Simulation study results for the orthogonalized age-period-cohort model fit with a PSS and RW2 prior model onto binomial data. The first and second rows are of the temporal effect and curvature for both the models alongside the true values. The bottom two rows are the bias and MSE box plots of the curvature for each model.

curve; and secondly, the curvature identification problem means the estimates for period and cohort will incur a larger penalty and differences between how each model smooths will be more noticeable. The smoothing parameter for the PSS model is considered 'optimal' since the cross-validation process used to estimate it is based on the data. This contrasts with the RW2 prior models which defines the smoothing parameter not with the data but with a prior. Due to this difference, the RW2 prior model may over smooth (depending on the prior) in comparison to the PSS model. This does not make the RW2 prior model inappropriate as the bias and MSE

boxplot show that the estimates are still remarkably close to the truth.

The direct comparison between the RW2 prior and PSS model estimates are in Figure B-3 in Appendix C. Comparisons between the full effects (the left-hand column) are not informative since these terms are unidentifiable, but comparisons between the curvatures (right-hand column) are. In each of the temporal curvatures, the estimates fall on the $y = x$ axis confirming the similarity between the estimates produced from each model. For both the period and cohort, where the cyclic pattern occurs, the estimates are condensed around zero. This is due to the attenuation from the Bayesian model and potential oversmoothing.

The similarity between the results in all three simulation studies give an empirical confirmation that the PSS model and the RW2 prior are both generating estimates extremely like one another. The results from RU and APC simulation studies confirm the RW2 prior model is appropriate for being fit to a reparameterised term as well as addressing the curvature identification problem. The slight oversmoothing of the cohort term and the general attenuation from the RW2 prior in the APC simulation study gives a clear indication that the RW2 prior is enforcing its own version of a penalisation against deviations in linearity. The main difference between the PSS and RW2 prior models is how each is defining the smoothing parameter. The PSS model uses the data to estimate an optimal smoothing parameter and the RW2 prior model uses a prior on the smoothing parameter.

3.5 Conclusion

We have drawn on the correspondence between penalised smoothing splines and random walk priors to compare the same model fit using `mgcv` and `r-inla`, respectively. The correspondence is based off the more general theory of reproducing kernel Hilbert spaces, but the existing literature may be inaccessible to practitioners. We provide an explanation that highlights the key reasons why there is an equivalence, and then move on to how we can use this equivalence in the context of APC modelling. A similar exposition of the relationship between penalised smoothing splines and stochastic partial differential equations was considered by (Miller et al., 2020) outside the context of APC modelling.

We define an identifiable APC model in terms of two of the three linear trends and three sets of curvature terms that are orthogonal to their respective linear trends (Holford, 1983). Based off the recommendation from Smith and Wakefield (2016), the RW2 priors are then placed on the curvature terms, and we show how the precision matrix of the RW2 prior is orthogonal to the addition of a constant and a linear trend. This reiterates the connection between the PSS model and the RW2 prior model since both penalty matrix and the precision matrix are rank deficient by two and their null spaces are spanned by the constant and linear functions.

To assess the correspondence between PSS models and RW2 prior models in general and understand if the identification problems in APC models cause this correspondence to falter, we performed three simulation studies. The first simulation study showed that both a PSS model and an RW2 prior model can be used interchangeably for smoothing. The second was for a univariate model that had been reparameterised into a linear slope and a set of orthogonal curvatures and gave no indication this process stopped the correspondence. The third and final simulation study was for a full APC model and the results showed that the RW2 priors are enforcing their own form of a penalty on deviations in linearity since the curvature identification problem was being alleviated. This not only provides evidence that the RW2 prior model is an appropriate for APC specific problems, but that a RW2 prior model is performing an equivalent job to a PSS model.

In addition to the simulation studies, a sensitivity analysis was performed on the choice of hyperprior used in the RW2 prior model. Whilst there was very little difference within the different specifications of PC or Gamma prior, there was a difference between using either a PC or a Gamma prior with each of the PC priors having a smaller MSE than the Gamma priors. Whilst the relative difference was noticeable, the absolute difference was still small.

An aspect we did not consider was how do the results change for non-constant unequal intervals; that is, when one of the time scales is of a non-constant different width to the others. For example, survey data from the Demographic and Health Surveys data used to model under-five mortality rates has age grouped into the following months $[0, 1)$, $[1, 12)$, $[12, 24)$, $[24, 36)$, $[36, 48)$ and $[48, 60)$ with yearly periods (USAID, 2019). From the work of Lindgren and Rue (2008), a RW2 prior model can still be fit to data at irregular intervals by using a Galerkin approximation to a weak solution of a stochastic partial differential equation. This is the default method of implementing a RW2 prior model in `r-inla`.

Outside of considering bias and MSE, we did not consider any other model validation procedures. For example, a leave-one-out cross validation is a powerful tool for assess a models predictive performance. Attaining predictions for both models is relatively straight forward, but prediction in a Bayesian paradigm, especially at new locations, is extremely intuitive. For APC modellers, predicting future changes in mortality in general and in relation to each temporal trends are an important component of evidence-based policy making.

3.6 Chapter conclusions

Since the correspondence between PSS models and RW2 prior models is based upon an area of maths that has little cross over with applied users of both models, it is fair to assume many would not be aware of it. To rectify that, we have given a brief overview of the main theoretical relations between the two and have provided a set of empirical results to show how this correspondence holds for different implementations of the models. For those who wish to use RW2 prior models to address the curvature identification problem, we showed how to define a reparameterised APC model with RW2 priors on the curvatures and gave robust evidence that the RW2 prior is indeed appropriate for fitting APC models to data unequally aggregated.

The results from Chapters 2 and 3 have provided two methods to appropriately address the curvature identification problem. Since these results are predominately based upon simulation studies, we wish to provide the reader with an example of how these results can be used in active research. Therefore, Chapter 4 considers a novel application of an APC model in under-five mortality rate (U5MR) modelling. Currently, methods to find estimates of U5MR include age and period and have data aggregated in unequal intervals. Due to this, cohort has not been included since it would be introducing both the structural link and curvature identification problems into the model. Now we have robust evidence for two separate implementations that include all three temporal trends in one model and that use data aggregated in unequal intervals, we can provide a novel extension to the field of U5MR and include cohort alongside age and period.

CHAPTER 4

ESTIMATING SUBNATIONAL UNDER-FIVE MORTALITY RATES USING A SPATIO-TEMPORAL AGE-PERIOD-COHORT MODEL

This chapter consists of a prepared manuscript that is in the process of being reviewed by co-authors. This manuscript contains a novel application of our age-period-cohort (APC) model to model under-five mortality rates (U5MR). The goal of this work is to develop the field of public health by providing clear and informative inference on an important area of research.

This work was performed in collaboration with Dr Theresa Smith (University of Bath), Professor Jon Wakefield (University of Washington) and Dr Johnny Paige (Norwegian University of Science and Technology). The work was conducted by and written up by Connor Gascoigne with Dr Smith providing expertise for APC related work and Professor Wakefield and Dr Paige providing expertise for the work relating to survey sampling, current methods for estimating U5MR and the data.

Including cohort in the estimation of U5MR has long been a goal with the most current models only using age and period (Mercer et al., 2015; Wakefield et al., 2019; Li et al., 2019, 2020) since the data comes in unequal intervals meaning there will be both the structural link and curvature identification problems present when cohort is included as described in Chapters Two and Three. With the work from this thesis, we can now include cohort into models for U5MR that is within the framework developed by my collaborators at the University of Washington. As part of this work, I attended the University of Washington's Summer Institutes on 'Spatial Statistics in Epidemiology and Public Health' and 'Small Area Estimation' in 2020, have been a regular attendee of Professor Wakefield's group meetings for the past two years and in November 2021, conducted a research visit to the University of Washington to work alongside my collaborators in person. This work was able to be conducted due to a grant won from the International Relations

Office, University of Bath.

The modelling and estimation of U5MR is of great importance to the United Nations (UN) due to the Sustainable Development Goal (SDG) 3.2: To reduce U5MR to at least as low as 25 per 1000 live births by the year 2030. For the UN to achieve SDG 3.2, they have identified low-and-middle income countries (LMIC) as high priority for any interventions as these countries tend to have a higher U5MR over higher income countries. Whilst acceptable estimates can be found at the national level using current methods, the UN requires estimates at a subnational level since this is where any interventions occur. When modelling data from LMICs on the subnational level, the sparsity of the data leads to overwhelming uncertainty in the U5MR estimates using current methods. Therefore, the immediate goal is to find subnational estimates of U5MR with less uncertainty.

The novel application of our APC model to U5MR extends the current literature as we are now able to include cohort without suffering from any of the identification issues. In this chapter, we fit an APC model that is reparameterised into linear and curvature terms with random walk of order two (RW2) priors on the curvatures. We show how the inclusion of cohort in the model leads to a better fitting and better predicting model than when cohort is omitted. The model estimates are validated by comparing the results against methods used in the current literature at a national level. In addition, we validate the predictions by performing a leave-one-out cross validation.

Abstract

Ascertaining subnational estimates of under-five mortality rates (U5MR) is a vital goal for the United Nations to reduce inequalities in health and well-being across the globe. This paper proposes a novel application of a spatio-temporal age-period-cohort model for U5MR where current methods do not consider cohort, only age and period. The data used is survey data from the Demographic and Health Surveys, a survey with a complex stratified cluster design that our method fully accounts for. We use a Bayesian hierarchical model with terms to smooth over both temporal and spatial components and perform inference using integrated-nested Laplace approximations. Our results show that the inclusion of cohort is necessary to produce more stable estimates with an acceptable level of uncertainty in situations where data is sparse. We validate our results by comparing against current methods at a national level.

4.1 Introduction

The United Nations (UN) estimate under-five mortality (U5MR) at a subnational level in part to assess their Sustainable Development Goals (SDGs) target 3.2, "By 2030, end preventable deaths of new-borns and children under five years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 deaths per 1,000 live births and under-five mortality to as least as low as 25 per 1,000 live births" (United Nations, 2019). Interventions normally take place at a local authority level; therefore, estimation, and prediction of U5MR at a subnational level are important goals of the UN to ensure the maximum impact and cost effectiveness of any future intervention. For example, the optimal intervention may differ region-to-region, so nationwide interventions are less efficient in terms of time, impact, and cost.

The U5MR from developed countries, such as those from Europe and Northern America, are not of great concern to the UN since these are traditionally much lower than those, say, from sub-Saharan Africa where children are 14 times more at risk of dying in the first five years of life (UN IGME, 2021). Due to this, the UN focus their attention on finding U5MR estimates for low-and-middle income countries (LMIC) to reach the SDG 3.2. It is common for LMICs to not have a full census data, but rather have survey data and in particular, survey data from the Demographic and Health Surveys (DHS) (USAID, 2019). The DHS has conducted more than 400 surveys in over 90 countries and has collected data on a national (Admin-0), region (Admin-1) and county level (Admin-2).

Small area estimation (SAE) techniques are used to estimate U5MR for geographical regions where there are little to no samples available. Here we briefly review the main techniques, but for a comprehensive review of SAE, see Wakefield et al. (2020). The starting point for spatial modelling of U5MR using SAE techniques is to find the weighted (direct) estimates (Horvitz and Thompson, 1952). Each individual will have their own weight which is proportional to the inclusion probability; consequently, they represent the sampling design, non-response and post-stratification/ranking. Direct estimates are considered the gold standard for large samples since they are design consistent when the weights are reliable and stable. This means, as the percentage of people who are included in the samples gets closer to the total population, the direct estimate will converge to the true population rate. However, when data is sparse, such as on a subnational level, the direct estimates suffer from a large amount of sampling variability, and the variance for these methods becomes unacceptably large.

Considering the problem of data sparsity, Fay and Herriot (1979) introduced a model that uses a transformation of the direct estimates to gain precision by smoothing the direct estimates over space using a random effects model. The Fay-Herriot approach captures the concept that, in general, we expect data points in time (and spatial location) more likely to be similar to one another than not. By incorporating this idea using a random effects model, the direct estimates

can 'borrow-strength' from neighbouring estimates to calculate a more accurate estimate with a smaller variance. We call these estimates the smooth direct estimates.

The direct and smooth direct estimates capture the complex survey design using design weights. Methods such as cluster level models do not use design weights but capture the complex design through other methods, i.e., including survey design specific covariates in the model or by using an overdispersion parameter. An example of cluster level is by Wakefield et al. (2019) which combines a discrete time survival analysis (DTSA) approach with space-time models to estimate U5MR which have been smoothed over both spatial location and time.

In the cluster level models of Wakefield et al. (2019), the age at death, the year of death and the spatial location of where the death occurred has been taken into consideration, but the cohort (year of birth) has not. This is the same for the direct and smooth direct estimates. A cohort effect is a measure of long-term exposures that those of a similar age encounter together as they move through life. As cohort captures latent trends, the influence of a cohort effect would not be apparent straight away but would be at a later stage in life. The smoother nature of a cohort trend in comparison to a year trend can lead to smoother, more stable predictions. For sparse data sets, such as those from the DHS on a subnational level, predictions will always be unstable and have a large amount of variation; hence, the inclusion of cohort can be of great importance as it would reduce the uncertainty for policymakers who rely on estimates of subnational U5MR to make vital decisions about interventions.

To model the U5MR at a subnational level with the inclusion of a covariate for (birth) cohort alongside those for age and period (year of death), we opt to use an age-period-cohort (APC) model. In U5MR, age is the most important temporal trend due to how much mortality changes in the earlier months; in developing countries, the first month of life can account for almost 50% of all under-five deaths (UN IGME, 2021). For example, in the 2014 Kenyan DHS (KDHS) (KDHS, 2015), deaths in the first month of life account for 42% of the total number of deaths. The period (year) effect is a measure of short-term exposures, such as a new treatment and the (birth) cohort effect is a measure of longer-term exposures. APC models have been used in several different health concerns such as suicide rates (Riebler et al., 2012b), stomach cancer incidence (Papoila et al., 2014), all-cause mortality (Smith, 2018) and opioid overdose mortality (Chernyavskiy et al., 2020). APC models have a well-known identification problem due to the linear dependence between the three temporal effects ($cohort = period - age$) but many authors have addressed this and there are several appropriate approaches (Holford, 1983; Clayton and Schifflers, 1987).

In this paper, we use an APC model to attain subnational estimates for U5MR using the 2014 KDHS data. Within the APC model, we acknowledge factors relating to the complex survey design by accounting for effects such as the stratum and clusters. The need for an APC model to find subnational estimates of U5MR is because current methods such as direct estimates

(Horvitz and Thompson, 1952), smooth direct estimates (Fay and Herriot, 1979) and other cluster level models (Wakefield et al., 2019) do not consider a cohort effect which could be vital in the reduction of uncertainty in the predictions. In Section 2, we describe the data used. In Section 3, we introduce the methodology for fitting an APC model and calculating the estimates of U5MR. Section 4 contains the results from the model at the national and subnational levels and a leave-one-out cross validation process. Section 5 concludes the paper.

4.2 Data

A typical DHS survey is a stratified two-stage cluster sampling scheme with stratification by county crossed with an urban/rural indicator drawn from an existing sample frame (generally the most recent census frame), which is a complete list of all sampling units that cover the target population. The first stage of sampling is to select the primary sampling units with probability proportional to size, these are called enumeration areas (EAs) and form the survey clusters. In the second stage, a fixed number (typically 25-30) of households are selected by equal probability systematic sampling from a list of all households in each EA.

We use the 2014 Kenyan DHS (KDHS) (KDHS, 2015) to produce an estimate for the U5MR for each of Kenya's 47 regions. The 2014 KDHS is stratified by the 47 regions each with its own urban/rural indicator. The total number of strata was 92 since both Nairobi and Mombasa are entirely urban. In the first sampling stage, 1,612 clusters were selected out of a possible 96,251 (as defined from the 2009 Kenya Population and Housing Census) of which 617 are urban and 995 are rural; urban areas are over sampled. In the second stage of sampling, 40,300 households are used from the selected EAs (25 households per EA).

Figure 4-1 shows a map of Kenya including region borders and cluster locations, reflecting the population density and the urban/rural stratification. For confidentiality, the GPS co-ordinates for cluster centres are jittered with urban and rural co-ordinates displaced by up to 2 and 5km, respectively. In addition, the locations for a further 1% of rural locations are displaced by up to 10km.

The data we shall use to estimate U5MR are from the DHS birth history questionnaire results. The birth histories questionnaire is answered by all aged 15-49 who spent the previous night in the sampled household and contains information on pregnancy, postnatal care, immunisation, and health of children born in the last five years. Specifically, we concern ourselves with answers relating to the children that give information on age at death (if occurred), period (year) of death (if occurred) and year of birth (cohort). In addition, the questionnaires contain the survey design information such as the cluster identification, region, and stratification. Each row of the survey relates to an individual child. We use yearly periods between 2006-2014 and monthly

ages between 0-60.

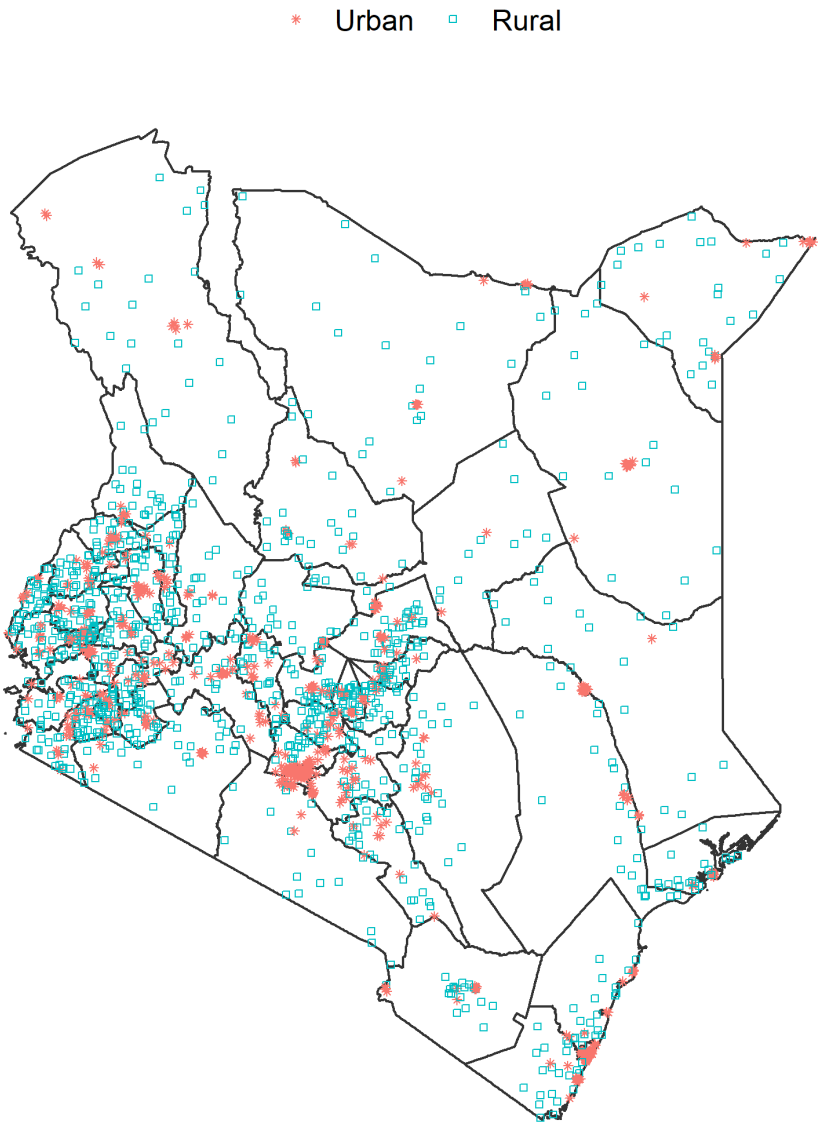


Figure 4-1: Urban and rural cluster locations on the 2014 Kenyan DHS with the 47 regions boundaries.

4.3 Methods

4.3.1 Discrete time survival analysis

We model under-five mortality rates (U5MR) using discrete time survival analysis (DTSA). DTSA allows for flexible modelling of (potentially) complicated temporal relationships; an important feature due to the substantially different hazards into the first five years of life (Clark et al., 2013).

To understand how DTSA is used to attain estimates of U5MR, first lets consider the case of a general DTSA. Let $m = 0, 1, \ldots, M$ be monthly ages where m^* is a particular month. The hazard H for month m^* is defined as the probability of event (in our case, death) occurring in the month m^*, given it has not occurred before month m^*

$$\mathrm{H}\left(m^*\right) = p\left(m = m^* | m \geq m^*\right).$$

The hazard can be thought of as the compliment to probability of surviving up to and including the given time. Letting $\text{Survival}\left(m^*\right)$ be the surviving probability beyond month m^*,

$$\begin{aligned}
\text{Survival}\left(m^*\right) &= p\left(m > m^*\right) \\
&= p\left(m > m^* | m \geq m^*\right) p\left(m > m^* - 1 | m \geq m^* - 1\right) \times \cdots \times p\left(m > 1 | m \geq 1\right) \\
&= \left[1 - \mathrm{H}\left(m^*\right)\right] \left[1 - \mathrm{H}\left(m^* - 1\right)\right] \times \cdots \times \left[1 - \mathrm{H}\left(1\right)\right].
\end{aligned} \tag{4.1}$$

where the total law of probability was used to split the survival beyond month m^* into conditional survivals. Each conditional is the survival probability for that group given they have survived up to the start of the group; it is the compliment of the hazard probability.

To calculate the U5MR, it is easier to consider $\text{Survival}\left(m^* = 59\right)$, the probability of surviving beyond the 59^{th} month. Therefore, subtracting this from one is the probability of not surviving beyond the 59^{th} month, the U5MR. Furthermore, following on from Mercer et al. (2015) and Wakefield et al. (2019), we group the months into six discrete hazards $[0, 1), [1, 12), [12, 24), [24, 36), [36, 48)$ and $[48, 60)$ where we attribute each month to one of the

hazards using

$$a\left[m\right]=\begin{cases}0.5 & \text{if } m=0,\\ 6 & \text{if } m=1,\ldots,11,\\ 17.5 & \text{if } m=12,\ldots,23,\\ 29.5 & \text{if } m=24,\ldots,35,\\ 41.5 & \text{if } m=36,\ldots,47,\\ 53.5 & \text{if } m=48,\ldots,59,\end{cases} \tag{4.2}$$

for $m=0,\ldots,59$ for each of the m months a child is at risk in. The midpoint of each of the age groups are used as this captures the inconsistency in the widths for the age intervals.

With each month attributed to one of the six discrete age groups, we define the U5MR as the compliment of the survival probability up to and including the 59^{th} month of life,

$$\begin{aligned}\text{U5MR} &= 1-\text{Survival}\left(m^{*}=59\right)\\ &= 1-\left(\left[1-\text{H}_{a[59]}\right]\left[1-\text{H}_{a[58]}\right]\times\cdots\times\left[1-\text{H}_{a[0]}\right]\right).\end{aligned} \tag{4.3}$$

where Eq.(4.1) is used between the first and second lines. A key assumption we use is that the hazard is constant within age groups, in other words

$$\begin{aligned}\text{H}_{a[48]}=\cdots=\text{H}_{a[59]}&=\text{H}_{52.5}\\ \text{H}_{a[36]}=\cdots=\text{H}_{a[47]}&=\text{H}_{41.5}\\ \text{H}_{a[24]}=\cdots=\text{H}_{a[35]}&=\text{H}_{29.5}\\ \text{H}_{a[12]}=\cdots=\text{H}_{a[23]}&=\text{H}_{17.5}\\ \text{H}_{a[1]}=\cdots=\text{H}_{a[11]}&=\text{H}_{6}\\ \text{H}_{a[0]}&=\text{H}_{0.5}.\end{aligned}$$

Therefore, the expression in Eq.(4.3) can be simplified,

$$\begin{aligned}\text{U5MR} &= 1-\left(\left[1-\text{H}_{52.5}\right]^{12}\left[1-\text{H}_{41.5}\right]^{12}\left[1-\text{H}_{29.5}\right]^{12}\left[1-\text{H}_{17.5}\right]^{12}\left[1-\text{H}_{6}\right]^{11}\left[1-\text{H}_{0.5}\right]\right)\\ &= 1-\prod_{a=1}^{6}\left[1-\text{H}_{a}\right]^{z[a]}\end{aligned} \tag{4.4}$$

where $a=0.5,6,17.5,29.5,41.5,52.5$ and $z\left[a\right]=1,11,12,12,12,12$ are the midpoints and number of grouped months in each of the six discrete age groups, respectively.

Now the discrete hazards are defined, for each child, define a Bernoulli random variable for the months lived to yield a sequence of binary outcomes 0/1 for survived/died which is of length 60 (these are also attributed to a cluster, region, strata, year, and cohort). In this format we have information on when each child survived/died as well as how many months at risk they

contributed to. For example, a child who dies after 15 months observed has contributed one, eleven and three months at risk to $[0, 1)$, $[1, 12)$ and $[12, 24)$, respectively, up to and including the month of death. This process is repeated for all individuals with the number of deaths and months at risk aggregated for each group.

Table 4.1 shows the first and last three rows of the child-month data. The columns from left-to-right are the cluster k, region r, strata (urban/rural), age group $a\,[m]$, period p, cohort c, total number of deaths for each group $y_{a[m],p,c,k}$ and total number of months at risk for each group $n_{a[m],p,c,k}$. From this point, if we refer to the KDHS data we are referring to the re-formatted version seen in Table 4.1.

Cluster	Region	Strata	Age	Period	Cohort	Deaths	Months at Risk
1	Nairobi	Urban	0	2006	2006	1	2
1	Nairobi	Urban	0	2007	2007	0	3
1	Nairobi	Urban	0	2008	2008	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1612	Busia	Urban	48-59	2013	2008	0	7
1612	Busia	Urban	48-59	2013	2009	0	22
1612	Busia	Urban	48-59	2014	2009	0	15

Table 4.1: Kenyan 2014 DHS data after being reformatted into child-months.

4.3.2 Spatio-temporal APC model

We use APC methods to model and analyse the temporal trends and SAE methods to explain variation in observations that can be attributed to geographical location and the effect of the survey design to find estimates for the U5MR. In this paper, we combine the APC ideas of Gascoigne and Smith (2021) with the SAE of Wakefield et al. (2019) to define an APC model with spatial components.

For the KDHS data, let $y_{a[m],p,c,k}$ be the number of observed deaths and $n_{a[m],p,c,k}$ be the number of monthly risks for age group $a\,[m]$, period p, cohort c and cluster k. As we use the true birth cohorts, there are instances where we have multiple cohorts for a single age and period combination (see rows 4 and 5 of Table 4.1), rather than just one if they were calculated using $c = p - a$; therefore, we explicitly denote c. However, we do not explicitly denote region r as this is contained within the cluster k. Including the three temporal effects together in one model causes the well-known identification issue due to the linear dependency $cohort = period - age$ (Clayton and Schifflers, 1987; Osmond and Gardner, 1989). To model the monthly probability of death conditional on the child being alive at the start of a month, consider the overdispersed

binomial, cluster-level model

$$y_{a[m],p,c,k}|\pi_{a[m],p,c,k}, d \sim \text{BetaBinomial}\left(n_{a[m],p,c,k}, \pi_{a[m],p,c,k}, d\right)$$

where $\pi_{a[m],p,c,k}$ is the monthly hazard for age group $a\,[m]$, period p, cohort c and cluster k. In a Beta-binomial model, d is the overdispersion (excess binomial variation) parameter. Overdispersion is common when modelling health and demographic data - especially when including effects for space and time. In data from the DHS, the cluster effects account for the clustering aspect of the survey design, and the dependence between respondents from the same households and dependence within the same cluster cause the overdispersion.

To reflect the over and under sampling of urban and rural areas common in the DHS survey data, we include a binary stratification variable for the urban/rural classification. This is necessary since oversampling of urban or rural clusters can lead to bias as the U5MR of urban areas may be very different to that of rural areas. Alongside the stratification, we include terms for the temporal, spatial and spatio-temporal effects.

Using the standard logit transform, the spatio-temporal APC model is

$$\text{logit}\left(\pi_{a[m],p,c,k}\right) = \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_{a[m]} + \eta_p + \xi_c + S_{r[s_k]} + \delta_{p,r[s_k]} \quad (4.5)$$

where $I\left(\boldsymbol{s}_K \in \text{ urban }\right) = 1$ if cluster k at location $\boldsymbol{s}_k$ is urban so that β_1 is the intercept for rural clusters and $\beta_1 + \beta_2$ is the intercept for urban clusters. The spatial term, which contains both the structured and unstructured spatial random effects, and the space-time random effects are denoted $S_{r[s_k]}$ and $\delta_{p,r[s_k]}$, respectively. When we say space-time, we are referring to space-period, but use space-time instead as this is in-line with the literature's naming convention. For any term relating to a region r, the notation $r\,[s_k]$ reads "the region r where cluster s_k resides".

In Eq.(4.5), we include a space-time interaction term to be consistent with the literature but there are other interaction terms that can be considered. For example, for spatio-temporal interactions there are the age-space and cohort-space interactions and for within temporal interactions there are the age-period, age-cohort, and period-cohort interaction terms. We chose not to include these now, but instead save this for future work.

Due to the linear dependence between the three temporal effects, the full age, period, and cohort terms are unidentifiable. To ensure identifiability, the APC terms are reparameterised into a set of identifiable terms (curvatures) and an arbitrary two (out of three) of the temporal linear trends (Holford, 1983). The curvatures are orthogonal to an intercept and a linear trend. The two region-specific, temporal linear trends are $t_{1,r}$ and $t_{2,r}$ and the curvatures are $\nu_{a[m]}$, η_p and ξ_c, for age, period, and cohort, respectively. We chose to retain the age and period linear trends (a cross-sectional model) since age is normally kept for its importance and the smaller range of years considered implies long term linear trends (such as cohort) might not be as important as

the short-term linear trends (period).

In addition to the identification issues due to the linear dependence between the APC temporal slopes, the curvature terms will suffer from the curvature identification problem (Gascoigne and Smith, 2021) due to the APC model being fit to data aggregated in unequal intervals. If the curvature identification problem is not addressed, the estimated period and cohort effects will display a saw-tooth pattern (Holford, 2006). The curvature identification problem is alleviated by including a penalty on the second derivative of the curvature terms, which in a Bayesian paradigm is achieved using a random walk two (RW2) prior (Rue and Held, 2005) for each of the curvature terms.

4.3.3 Bayesian Inference

We fit the model using a Bayesian hierarchical model in which prior distributions need to be assigned to all the random effect parameters in the model. To alleviate the curvature identification problem, we use RW2 priors for each of the curvature terms. That is, $\nu_{a[m]}|\tau_\nu \sim \text{RW2}(\tau_\nu)$, $\eta_p|\tau_\eta \sim \text{RW2}(\tau_\eta)$ and $\xi_c|\tau_\xi \sim \text{RW2}(\tau_\xi)$ where τ_ν, τ_η and τ_ξ are the precision parameters for the age, period, and cohort curvature, respectively. RW2 priors are commonly defined for uniformly spaced time steps, which is not the case for age in the KDHS. It is possible to attain estimates using a RW2 prior at irregular time steps by using a Galerkin approximation to the solution of a stochastic differential equation (Lindgren and Rue, 2008).

In the spatial term, $S_{r[s_k]}$, there is a structured and unstructured spatial component. Independently, the structured part can be modelled using an intrinsic conditional autoregressive (ICAR) model (Besag et al., 1991), and the unstructured part can be modelled with an independent identically distributed model. We choose to model them together using a BYM2 model (Riebler et al., 2016), $S_{r[s_k]}|\tau_s, \phi \sim \text{BYM2}(\tau_S, \phi)$, where $\phi \in [0, 1]$ is the mixing parameter that measures the proportion of the marginal variance, τ_S^{-1}, that is explained by the structured spatial effect.

Since we believe the period trend will be different from region-to-region whilst also being structured in space, we assume a Type IV space-time interaction for $\delta_{p,r[s_k]}$ (Knorr-Held, 2000). As with the main period and spatial effects, we assume a RW2 and ICAR prior distribution, respectfully.

For all distributions we use penalised complexity (PC) priors (Simpson et al., 2017) on precision and correlation components such as τ and ϕ. A PC prior for a given model component is defined $\text{P}(\text{model parameters } > U) = p$ where U is a sensible upper bound and p is the probability of the model parameter being in the upper bound. We follow Li et al. (2020) for the PC priors hyperparameters specification for each model term. The precisions for all temporal curvatures have $U = 1$ and $p = 0.01$. For the BYM2 term, the variance and scale parameters have $U = 1$

and $p = 0.01$ and $U = 1/2$ and $p = 2/3$, respectively. Finally, the space-time interaction term has $U = 1/2$ and $p = 2/3$ for the joint space-time components. More details on the priors are in Appendix C.

4.3.4 Implementation

We fit the spatio-temporal APC model with integrated nested Laplace approximations (INLA) as implemented in the `r-inla` package Rue et al. (2009). INLA provides accurate approximations of the marginal posterior distribution for random effects and hyper-parameters whilst avoiding the need for costly and time-consuming Markov-chain Monte Carlo (MCMC) sampling.

The RW2 and BYM2 components are directly available from `r-inla` using `rw2` and `bym2` options of the `model` argument of `f(...)`. The default method to implementing a RW2 model in `r-inla` is by the Galerkin method of Lindgren and Rue (2008); consequently, no additional work is needed to include the irregular interval age term. The space-time component is not directly available and instead needs to be specified through a `generic0` latent model alongside a constraint (precision) matrix, `Cmatrix`, and any additional constraints, `extraconstr`. An explanation of how to implement the space-time interaction can be found in Appendix C.

We can sample from the monthly hazard posterior distribution (PD) using `r-inla`, but are not able to sample from the U5MR PD directly as the formula for U5MR requires non-linear combinations of samples from the PD. To generate a sample for the U5MR PD, we first sample from the monthly hazards PD using `r-inla`, and then use the formula in Eq.(4.6) to define a U5MR PD. As we do not include any other stochastic processes when defining the U5MR PD, this is an acceptable method. More details are in Appendix C.

The full code for implementing the APC (and any other) model considered in this paper can be found at `https://github.com/connorgascoigne/APC-models-for-U5MR`

4.3.5 Estimation

We define the U5MR for period p and region r using

$$\text{U5MR}_{p,r} = 1 - \prod_{a=1}^{6} \left[1 - \text{expit}\left(\text{logit}\left[\pi_{a,p,c,r}\right]\right)\right]^{z[a]} = 1 - \prod_{a=1}^{6} \left[\frac{1}{1 + \exp\left(\text{logit}\left[\pi_{a,p,c,r}\right]\right)}\right]^{z[a]} \tag{4.6}$$

where $a = 0.5, 6, 17.5, 29.5, 41.5, 52.5$ and $z\left[a\right] = 1, 11, 12, 12, 12, 12$ are the midpoints and number of grouped months in each of the six discrete age groups, respectively. This formula is an adaption of the U5MR formula, Eq.(4.4), discussed previously.

Given we are fitting a stratified cluster level model, we cannot directly find the regional U5MR. Instead, we find an U5MR estimate for the stratified region and then aggregate over the stratification to find the regional U5MR (Paige et al., 2022). The U5MR for the urban and rural areas in period p for region r are

$$\text{U5MR}_{p,r,\text{rural}} = 1 - \prod_{a=1}^{6} \left[\frac{1}{1 + \exp\left(\beta_1 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_a + \eta_p + \xi_c + S_r + \delta_{p,r}\right)} \right]^{z[a]},$$

$$\text{U5MR}_{p,r,\text{urban}} = 1 - \prod_{a=1}^{6} \left[\frac{1}{1 + \exp\left(\beta_1 + \beta_2 + t_{1,r}\beta_3 + t_{2,r}\beta_4 + \nu_a + \eta_p + \xi_c + S_r + \delta_{p,r}\right)} \right]^{z[a]}.$$

The aggregated U5MR for period p and region r is

$$\text{U5MR}_{p,r} = \text{U5MR}_{p,r,\text{rural}} \times q_{p,r} + \text{U5MR}_{p,r,\text{urban}} \times (1 - q_{p,r}) \tag{4.7}$$

where $q_{p,r}$ is the proportion of the target population in period p and region r that is rural. We use the true proportions for each period-region combination to account for urbanisation and the over and under sampling of urban and rural areas, respectively. An estimate for the national U5MR can be defined by multiplying $\text{U5MR}_{p,r}$ by the yearly national proportions of each region. The proportions in question are found using the method developed by my collaborators at the University of Washington which is yet to be published. Examples of the strata and national proportions can be found in Figures C-1 and C-2 in Appendix C.

For a full U5MR PD for region r and period p, we extract the full PD of the monthly hazards from `r-inla`, and then use the formulas of this section to define the U5MR PD.

4.4 Results

Alongside the results for the APC model, we include the results from an age-period (AP) and age-cohort (AC) model. Both the AP and AC models do not have the structural link between the temporal terms so are not overparameterised. However, we still reparameterise them in terms of a linear slope and their orthogonal curvatures, but no longer drop any of the slopes. The AP and AC models are

$$\text{AP:} \quad \text{logit}\left(\pi_{a[m],p,c,k}\right) = \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + a\left[m\right]_r\beta_3 + p_r\beta_4 + \nu_{a[m]} + \eta_p + S_{r[s_k]} + \delta_{p,r[s_k]}$$

$$\text{AC:} \quad \text{logit}\left(\pi_{a[m],p,c,k}\right) = \beta_1 + I\left(\boldsymbol{s}_k \in \text{ urban }\right)\beta_2 + a\left[m\right]_r\beta_3 + c_r\beta_4 + \nu_{a[m]} + \xi_c + S_{r[s_k]} + \delta_{p,r[s_k]}.$$

Besides the AP and AC models, the other nested models (within the APC model hierarchy) include the univariate temporal models or do not include age. The other nested models are not included since the current literature include age and period at a minimum when modelling

U5MR. The AP model is equivalent to the cluster level beta-binomial model used in practise (Li et al., 2020), which we shall call the age-time (AT) model. In a pre-analysis, we compared the AP model to the AT model from the `SUMMER` package in `R` (Li et al., 2020). As this is a comparison for the temporal terms only, we fit the AP and AT models without the spatial or spatio-temporal terms. The results are in Figures C-3 - C-4 in Appendix C. The slight difference between the two models is due to the AT model from `SUMMER` including an iid random effect for time (period) and the AP model not; hence, we say these models are equivalent rather than equal. However, the similarity between the results from the AP and AT models confirm the reparameterisation is not having any adverse effects and the model is working as expected.

We use the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and Watanabe-Akaike information criterion (WAIC) (Watanabe and Opper, 2010), which provide measures of model fit and model complexity, to compare between the APC, AP, and AC models. Table 4.2 shows the scores for each model. The best and worst scores are acknowledged by bold and italic fonts, respectively. The APC model has both the lowest scores; whereas, the AC model has both the highest scores, indicating the APC model is the best fitting, and the AC model is the worst fitting. In a pre-analysis, we considered each of the APC, AP and AC model with a different interaction type and found the Type IV interaction scored the best for both the APC and AC models and the Type II scored the best for the AP model. The APC model with a Type IV interaction scored the best overall. The relative difference between the scores when considering different interactions for the same model was far less than when looking across models with the same interaction. Consequently, the choice between APC, AP and AC models is far more influential than the specification of the space-time interaction model.

Test	APC	AP	AC
DIC	**17266.89**	17338.53	*17342.01*
WAIC	**17267.38**	17338.56	*17342.34*

Table 4.2: Model scores for each of an APC, AP, and AC models. The best entries are in **bold**, whilst the worst entries are in *italics*.

Table 4.3 shows the posterior summaries of all parameters for each of the APC, AP, and AC models. For each model, the relevant slope parameters are not significant, indicating the temporal linear trends are not influential in U5MR. The precision for each of the temporal curvatures are all large, indicating the concentration of the posterior distribution about the mean for each RW2. Similar statements can be said of the precision for the spatial and spatio-temporal effects across all three models. The spatial mixing parameter being closer to zero indicates most of the spatial dependence is unstructured, and this is similar across all three models.

	APC			AP			AC		
Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
Strata	-0.107	-0.003	0.100	-0.105	-0.001	0.102	-0.103	0.000	0.104
Age Slope	-4.283	-0.031	4.218	-4.327	-0.075	4.173	-4.329	-0.076	4.172
Period Slope	-0.056	-0.003	0.049	-0.055	-0.003	0.050	-	-	-
Cohort Slope	-	-	-	-	-	-	-0.055	-0.003	0.050
BetaBinomial Overdispersion, d	0.000	0.001	0.002	0.000	0.001	0.002	0.000	0.001	0.002
Age RW2 Precision, τ_ν	29.013	150.467	567.367	69.545	193.789	646.560	27.179	146.641	530.378
Period RW2 Precision, τ_η	14.912	165.977	1968.495	85.845	1268.967	105450.308	-	-	-
Cohort RW2 Precision, τ_ξ	59.853	784.568	19873.478	-	-	-	95.271	4928.381	210282.624
Region Precision, τ_S	4.829	7.923	12.924	4.876	7.997	13.099	4.836	7.723	12.441
Region Mixing, ϕ	0.011	0.102	0.518	0.006	0.080	0.377	0.013	0.100	0.500
Space-Time Precision, τ_δ	53.716	478.488	8747.622	118.408	616.269	9347.391	426.661	4177.737	320582.844

Table 4.3: Posterior quantile summaries for APC, AP, and AC models with a Type IV space-time interaction term.

4.4.1 National estimates for U5MR

We compared the aggregated subnational U5MR estimates from the APC, AP, and AC models against a national weighted (direct) estimate (Horvitz and Thompson, 1952) and a national smoothed (smooth) direct estimate (Fay and Herriot, 1979). Direct estimates are design consistent, but when data is sparse, they have a large amount of uncertainty. Smooth direct estimates are a transformation of the direct estimates modelled via a random effects model, which reduces the uncertainty, gaining precision, by smoothing over temporal trends.

The APC, AP, and AC aggregated U5MR estimates alongside the national direct and national smooth direct U5MR estimates are in Figure 4-2. In Figure 4-2, the APC and AP models follow the direct and smooth direct estimates well, often at times being between them. This contrasts with the AC model which provides a flatter curve that is still following the general trend. The flatness of the AC model reflects the fact that the cohort effect is a measure of long-term exposures and hence would be influenced less by the short-term yearly exposures. The 95% CI, displayed in Figure 4-3, show the APC, AP and AC estimates are always narrower than the 95% CI for both the direct and smoothed direct estimates, whilst also being contained within them for the most part.

The upturn in the direct, smooth direct, APC and AP estimates for the year 2014 seem unrealistic and does not follow the reducing trend we expect. This could be due to the 2014 KDHS being conducted in said year, reducing the total number of months at risk seen. The total number of months at risk for the year 2014 was 147,980, and for the years 2006 - 2013, the minimum, median and maximum number of months at risk were 222,484, 247,456 and 252,081, respectively. The lack of an upturn in the AC model is an indication that by including cohort, we have more stability in the estimates as it is less influenced by the short-term differences.

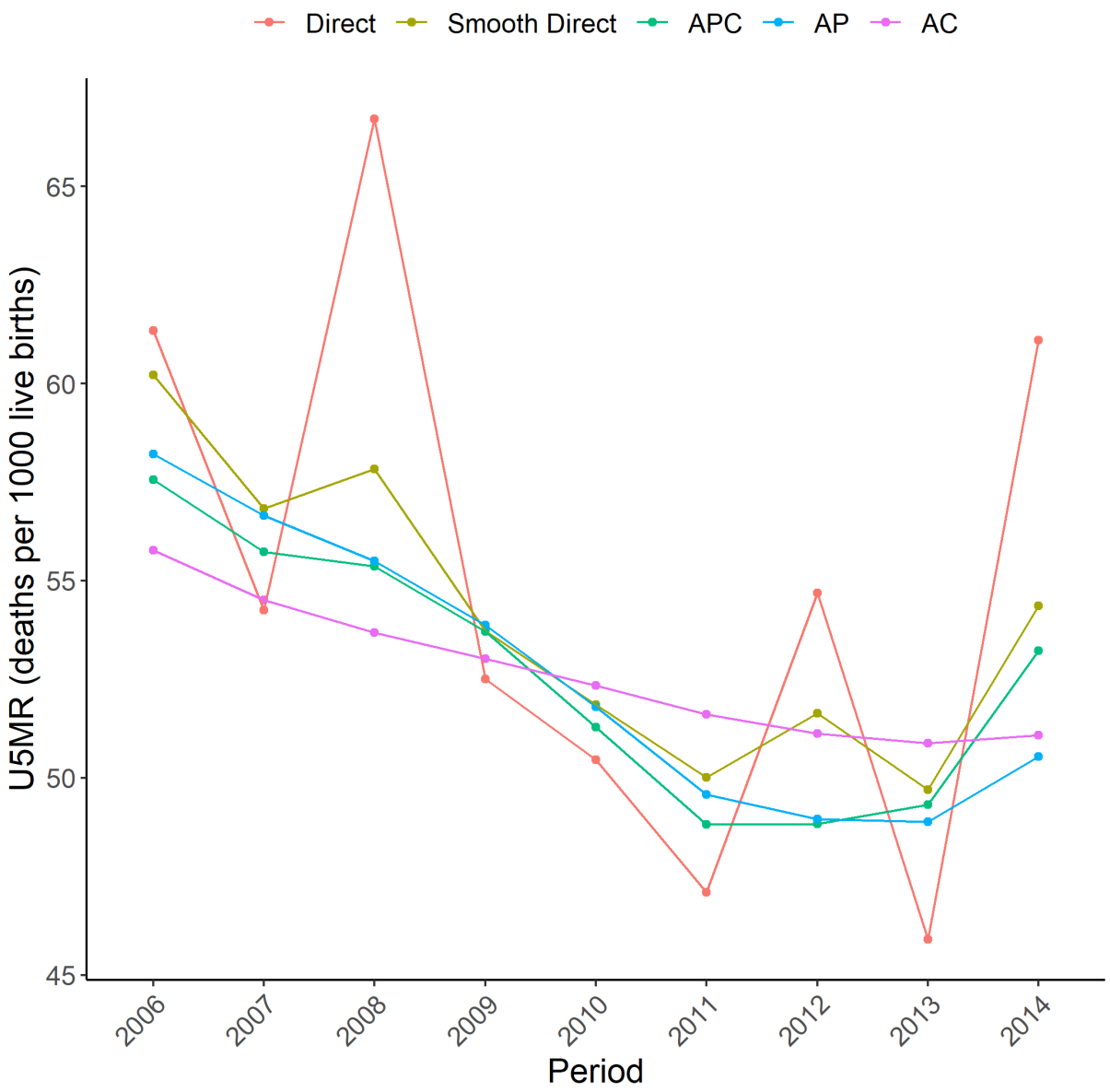


Figure 4-2: APC, AP and AC, direct and smooth direct national estimates of U5MR for Kenya, 2006-2014.

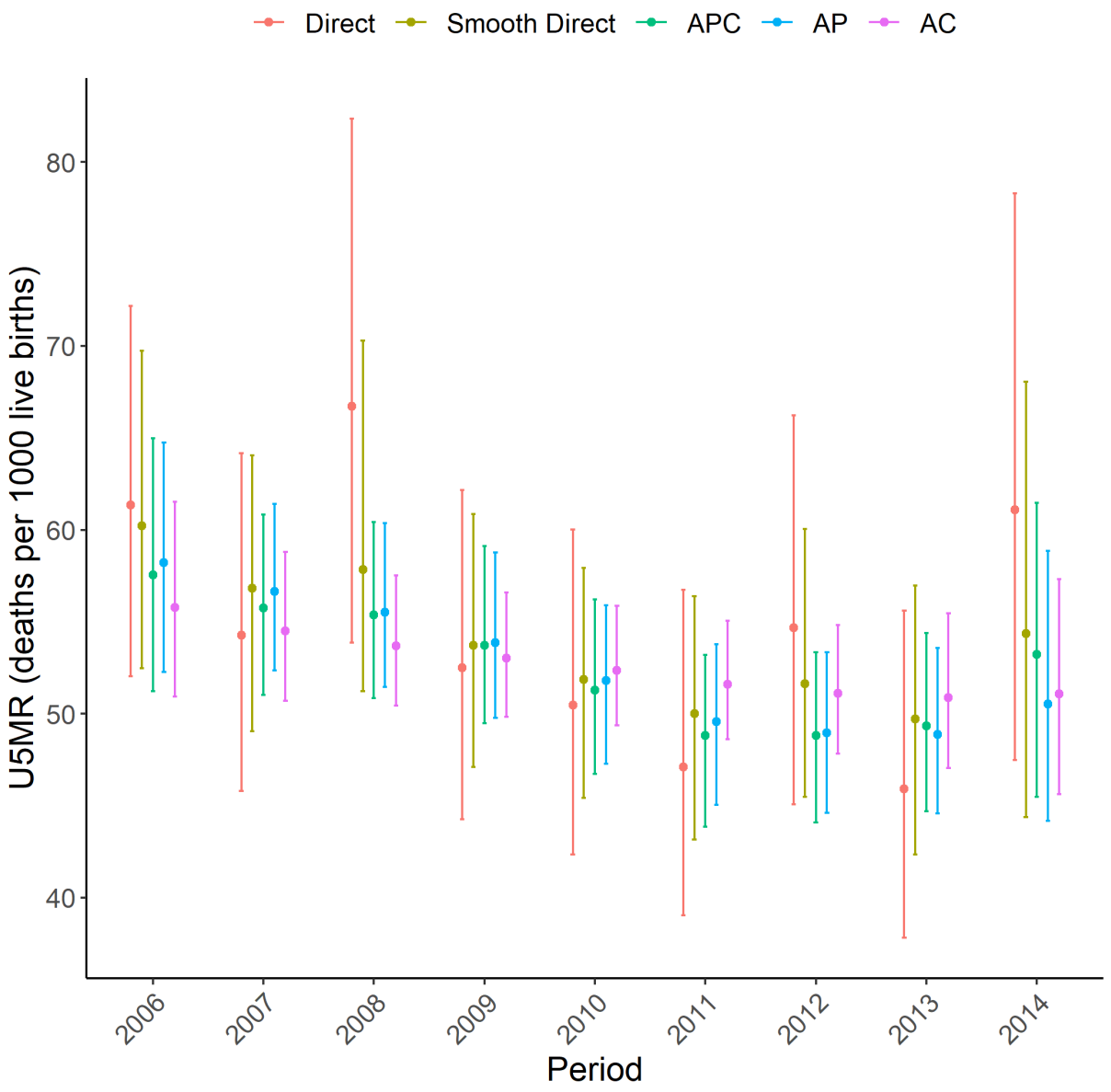


Figure 4-3: APC, AP and AC, direct and smooth direct national estimates of U5MR for Kenya, 2006-2014 including 95% credible intervals.

4.4.2 Subnational estimates for U5MR

The subnational results for the APC model are in Figure 4-4 and Figure 4-5 where the former is the median U5MR estimates per 1000 live births and the latter is the width of the associated 95% CI. In each Figure, moving from left-to-right, top-to-bottom are the years 2006-2014. In Figure 4-4, we see the spatial structure in the U5MR estimates as closer regions have similar rates. For example, regions in the west have a high U5MR whereas regions in the south have a lower U5MR. The influence of the space-time trend is hard to see from Figure 4-4 alone. To see the influence clearer, we include a line plot of the U5MR per 1000 live births for the individual regions, Figures C-5 - C-8 in Appendix C. If there is no space-time interaction, each of the lines would be parallel to one another; whereas, in the presence of a space-time interaction, the lines would cross. Since there are multiple lines crossing in each figure, there is indeed a space-time interaction.

Comparisons between Figure 4-4 and Figure 4-5 reveal that areas with a higher U5MR correspond to areas with higher uncertainty. This is due to the binomial likelihood where the mean and variance are proportional to one another. As expected, the width of the CI is reducing year-on-year except for 2014 which, as previously mentioned, is the year the survey is being conducted so is incomplete.

Figure 4-6 shows comparisons between direct, AP and AC estimates against the APC estimates for U5MR on the logit scale for each of the 47 regions over the years 2006-2014. The plots show signs of attenuation in the APC, AP, and AC estimates, which is due to the shrinkage, and is expected. The APC and AP model estimates are more like one another than the APC and AC model estimates, as expected given the national results from Section 4.4.1.

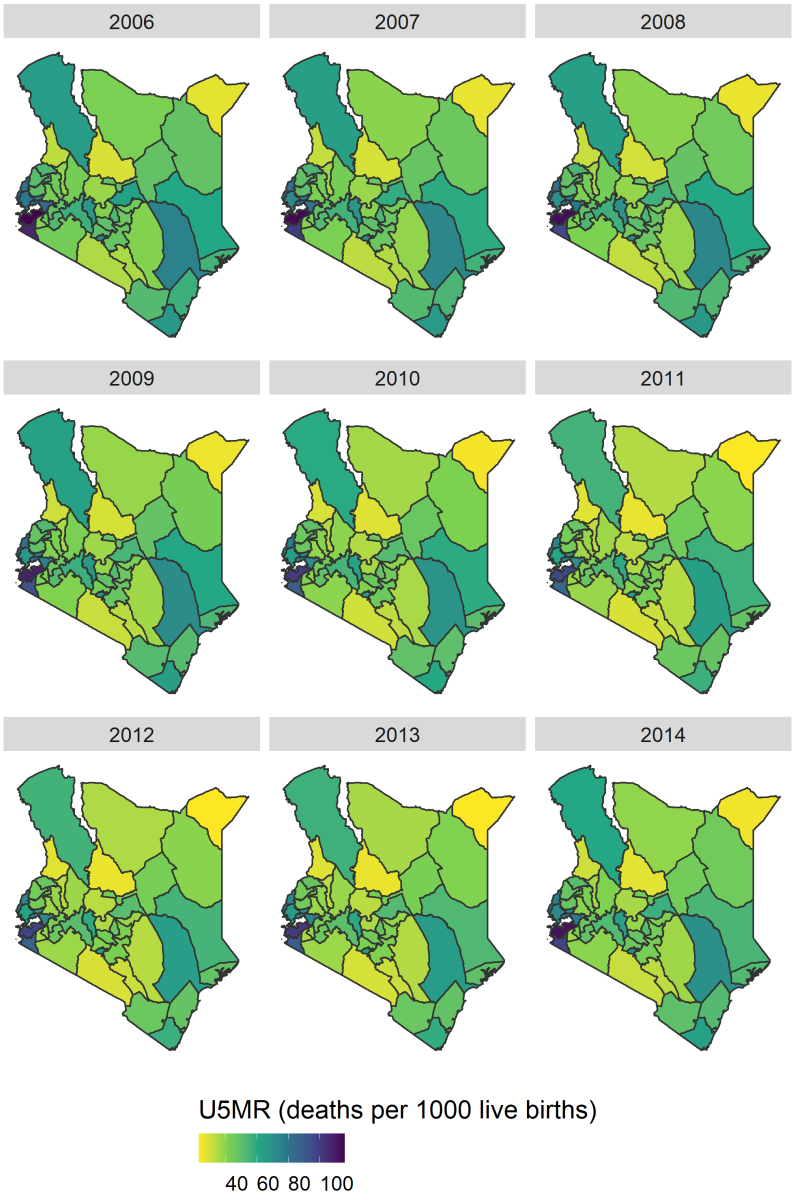


Figure 4-4: APC subnational estimates of U5MR for Kenya, 2006-2014.

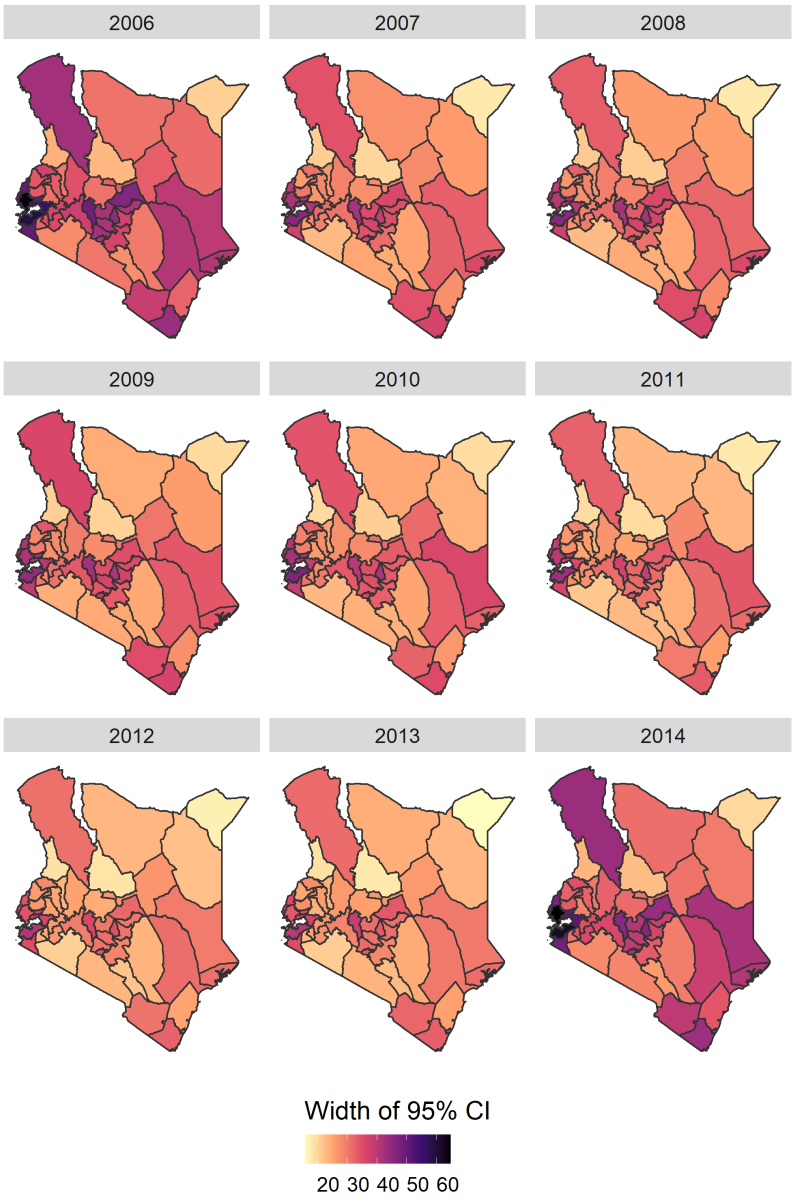
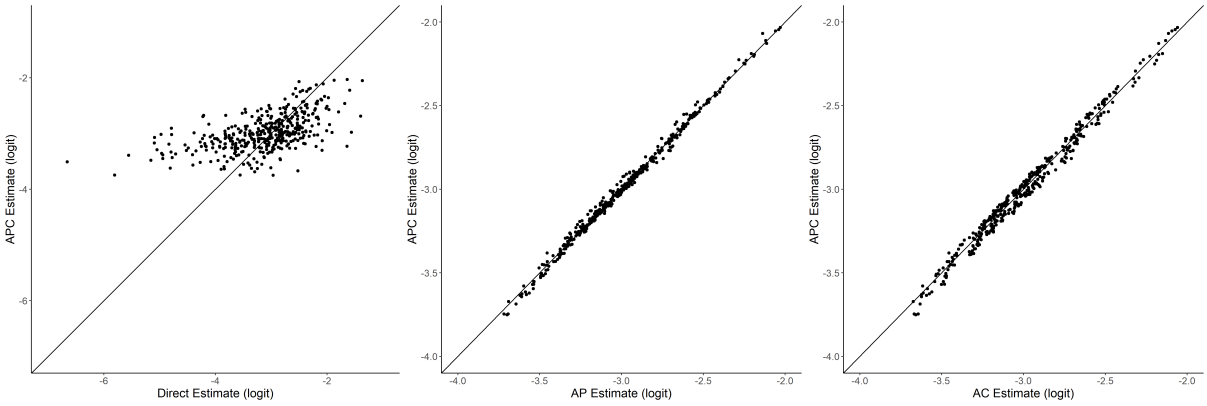


Figure 4-5: Width of 95% credible intervals. Width of the 95% credible intervals for the APC subnational estimates of U5MR for Kenya, 2006-2014.



(a) Versus direct estimates. (b) Versus AP estimates. (c) Versus AC estimates.

Figure 4-6: Yearly, regional direct, AP and AC estimates verses APC estimates on the logit scale. From left-to-right are the direct, AP and AC models, respectively.

4.4.3 Validation

The purpose of this paper is to produce a set of smooth estimates for U5MR with less variability when data is sparse. We have chosen to overcome the issue of data sparseness by smoothing across spatial and temporal trends with the novel inclusion of smoothing across cohort. The results of the previous section show that each of the APC, AP, and AC model produce estimates that are smoother with less variability than the direct estimates and of the three, the APC model was shown to be the best fitting.

Another way we can assess the suitability of the inclusion of cohort is to compare predictions against the actual direct estimates using a leave-one-out cross-validation (LOOCV). In this LOOCV, we systematically leave out any observations from the final period (2014) for each region and fit our APC, AP, and AC models to this "truncated" dataset to evaluate the predictive performance of our model. We use the direct estimates as the comparative value since they are design consistent even if they suffer from large sampling variability. We then compare the model predictions for each region against their respective direct estimates when fit to the full dataset.

Using `r-inla`, we produce a U5MR PD using the method in Appendix C but with the truncated dataset, rather than the full data. The U5MR PD will contain the sample variability of `r-inla`, but will not contain the (complex) design variability of the survey. To include this, and make the model estimates more comparable to the direct estimates, we use the sampling distribution (SD) of U5MR,

$$\widetilde{Y}_r^{(w),(m)} \sim \mathrm{N}\left(Y_r^{(w)}, \hat{V}_{\mathrm{r}}^{\mathrm{Des}}\right).$$

Here, $Y_{r,p} = \mathrm{logit}\left(\mathrm{U5MR}_{r,p}\right)$ which is different from $y_{a[m],p,c,k}$, the number of observed deaths. Furthermore, for simplicity, have dropped p from the subscripts since $p = 2014$ is fixed. The term $\widetilde{Y}_r^{(w),(m)}$ is the m^{th} draw from the SD and $Y_r^{(w)}$ is the w^{th} draw from the PD. We use this SD since it has been shown to perform well in the context of small area estimate for complex survey data (Mercer et al., 2014). The variability from `r-inla` is included in the draws from the PD and the variability from the complex design is included in $\hat{V}_{\mathrm{r}}^{\mathrm{Des}}$. More details on how we calculate the SD from the PD are in Appendix C.

To assess the comparability between the model estimates and the direct estimates, we use: difference-squared, variance, mean squared difference (MSD) and coverage. We use the term difference and MSD rather than bias and mean squared error since we are comparing to another estimate and not the truth. On the logit scale, we find the difference-squared, variance, MSD and coverage for each region and then average over all regions to give a final score. The scores

are calculated,

$$\begin{aligned}
\text{Difference}^2 &= \frac{1}{R}\sum_{r=1}^{R}\left(\overline{\widetilde{Y}_r} - Y_r\right)^2; \qquad \overline{\widetilde{Y}_r} = \frac{1}{M\times W}\sum_{m=1}^{M}\sum_{w=1}^{W}\widetilde{Y}_r^{(w),(m)} \\
\text{Variance} &= \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{(M-1)\times(W-1)}\sum_{m=1}^{M}\sum_{w=1}^{W}\left(\widetilde{Y}_r^{(w),(m)} - Y_r\right)^2\right] \\
\text{MSD} &= \frac{1}{R}\sum_{r=1}^{R}\left[\text{Difference}_r^2 + \text{Variance}_r\right] \\
\text{Coverage} &= \frac{1}{R}\sum_{r=1}^{R} I\left(Y_r \in 95\%\text{ CI for region } r\right).
\end{aligned}$$

where, Y_r is the direct estimate of the U5MR for region r in 2014 on the logit scale.

For the year 2014, there were two regions, Makeueni and Samburu, where there were no observations. Due to this, there are no direct estimates of U5MR for use in the difference-squared, variance, MSD, and coverage; therefore, these regions are omitted. The results of the LOOCV are in Table 4.4 where the best entry is in bold, and the worst entry is in italics; for all but the coverage, smaller values are favourable. Figure 4-7 shows the regional variation in scores for each of the APC, AP, and AC scores. Each model has its own colour, and the horizontal lines are the average of the 45 regions. Of the three, the APC model has the lowest variance and MSD, indicating it is closer to the direct estimate in average. Interestingly, each of the models have the same coverage and the regions that are missed are the same for each model: Mandera, Migori, Nairobi, Nyeri and Wajir.

Score	APC	AP	AC
Difference^2	*0.64*	0.63	**0.63**
Variance	**9.86**	*10.17*	10.09
MSD	**10.50**	*10.80*	10.72
Coverage (%)	**88.89**	**88.89**	**88.89**

Table 4.4: Validation scores for the leave-one-out cross-validation. The worst entry for each model when compared to the direct estimate is in *italics* whereas the best entry is in **bold**.

The variance and MSD for each of the models is slightly larger than we would like to see, but this is not necessarily indicative of a poorly fitting model. Due to the high sampling variability the subnational direct estimates suffer, we cannot use these as a 'true' U5MR, but rather an estimate with different properties to those from the APC, AP and AC models. Therefore, if the user wishes to have an estimate with design consistency but suffering from high sampling variability, they will choose the direct estimate. Alternatively, is the user wished to have an estimate that is smoother with less variance but not design consistent, they would choose the APC model as this is the most comparable to the direct estimate.

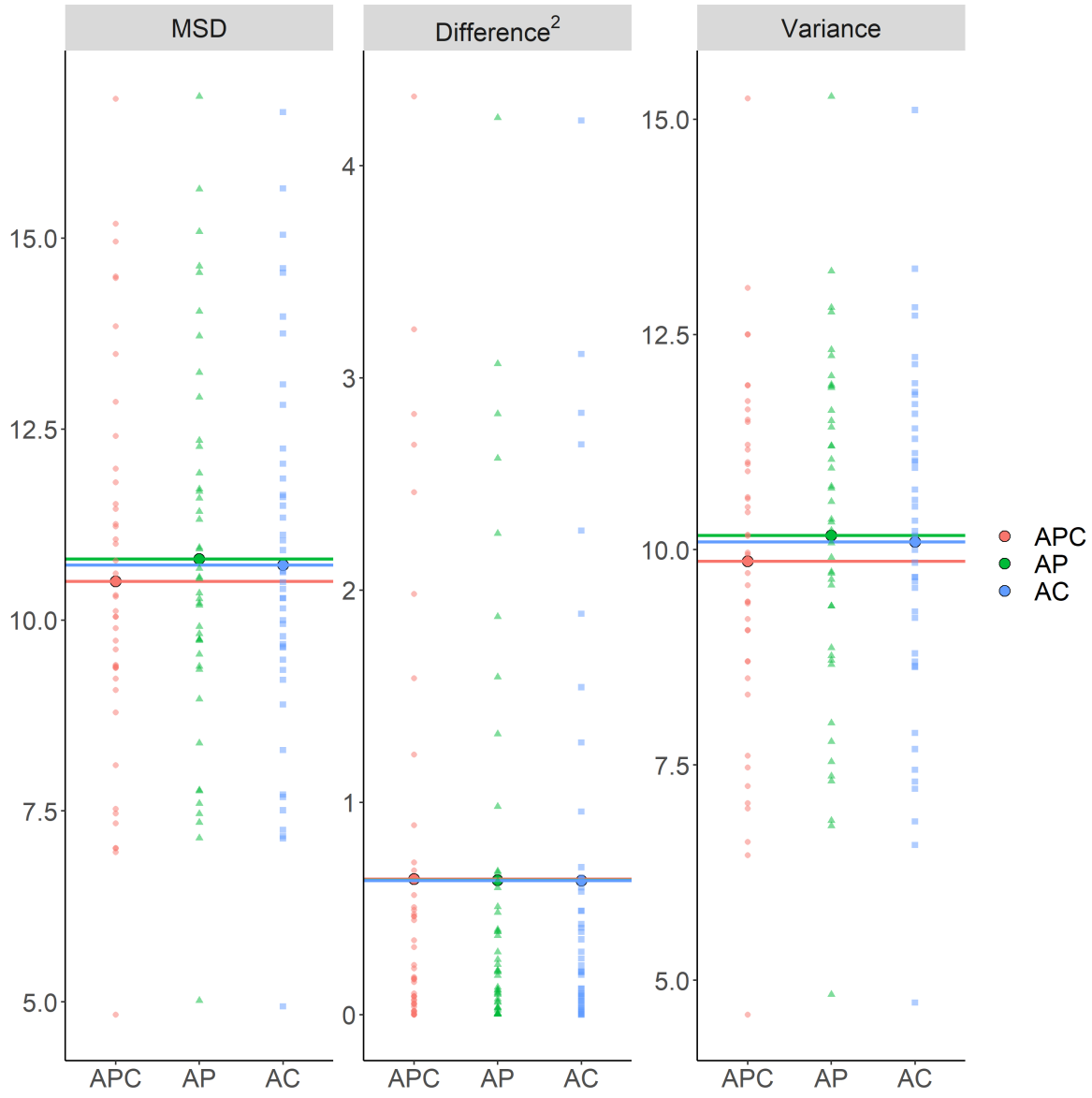


Figure 4-7: Regional variation for each of the validation scores.

A final way to understand the mean and variance for each region is via a ridge plot as it is a clear way to display the mean and variance for each of the U5MR for each region. Figure 4-8 is a ridge plot of the U5MR PD for the LOOCV. Along the x-axis in Figure 4-8 is the predicted deaths per 1000 live births for each of the regions in 2014, and along the y-axis are the regions in alphabetical order. The PD has a ridge for each of the regions in Kenya, since the SD is relying on the direct estimate variance, it does not have a ridge for the two missing regions. Therefore, we use the PD rather than the SD in Figure 4-8. This is also true for Figures C-11 and C-12 in Appendix C. The former is the median predicted values of U5MR per 1000 live births and the latter is the width of the 95% CI of the predictions. For the areas where there is high U5MR, the

CI width is wider and for the areas of low U5MR, the width is smaller. The spatial structure is still preserved for the predictions, in that neighbouring regions tend to have more similar U5MR than regions far apart. If we were to reproduce each of ridge and map figures using the SD, not the PD, we would expect to see a larger variance in the ridge and width plots but not too much difference in the median plot.

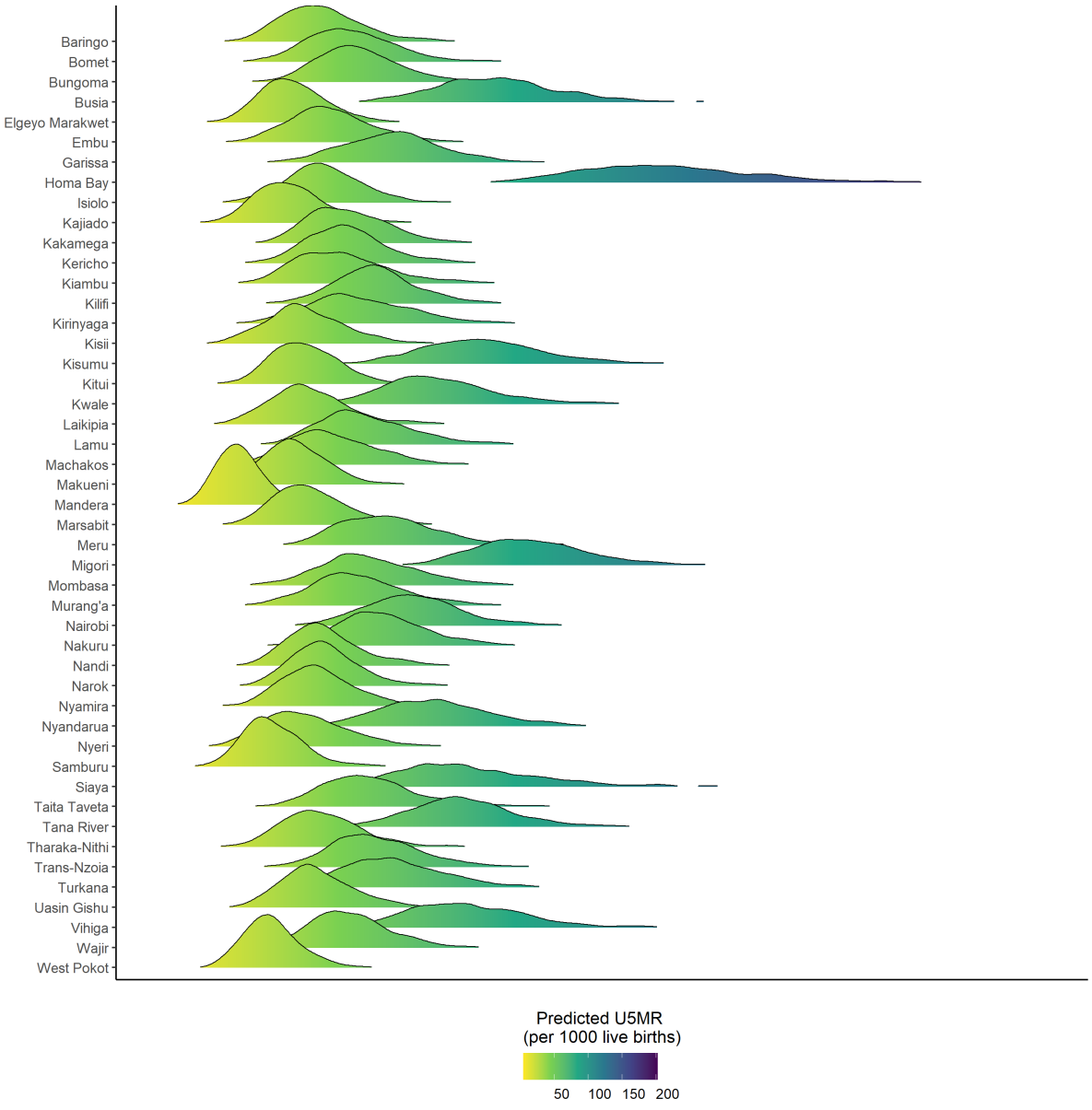


Figure 4-8: APC ridge plot of the posterior distribution for each of the 47 regions in 2014 from the leave-one-out cross-validation.

4.5 Conclusions

In this paper, we fit a spatio-temporal age-period-cohort (APC) model to the Kenyan 2014 Demographic and Health Surveys (DHS) survey data to produce subnational estimates for under-five mortality rates (U5MR). The inclusion of cohort alongside age and period in the context of U5MR is novel as previous methods only use age and period temporal trends. In addition, we fit age-period (AP) and age-cohort (AC) models (subclass models from the APC modelling hierarchy) to assess whether the inclusion of all three temporal trends is suitable or not. In addition to comparisons between each of the APC, AP, and AC models, we compared the results against weighted (direct) estimates (Horvitz and Thompson, 1952) and smooth direct estimates (Fay and Herriot, 1979) using both quantitative and qualitative methods.

Direct estimates are known to be the gold standard when data is plentiful since they are design consistent. However, when data is sparse, such as at the subnational level, the direct estimates suffer from a large amount of sampling variability. As an alternative to direct estimates, there are: smoothed direct estimates which models a transform of the direct estimates to smooth across time; and other cluster level models that make use of age and period temporal effects and smooth over both time and space (Wakefield et al., 2019) and each can be implemented using the `R` package `SUMMER` (Li et al., 2020). The theme across alternative methods is to include a form of smoothing across time and/or space to reduce the variability in the estimates. Of the models used in practise now, none of them smooth over cohort. The (birth) cohort is readily available in the DHS survey but is often overlooked with current methods favouring the more important age and period temporal trends since including cohort alongside age and period leads to problems relating to lack of identifiability of the temporal trends (Holford, 2006).

Predictions of U5MR are vital to advise on policy and interventions to meet the United Nations Sustainable Development Goal (SDG) 3.2 for each country to have a U5MR of 25 per 1000 live births by 2030. In particular, the production of estimates and predictions at a subnational level are the main objective since this is where any intervention aimed at achieving the SDG 3.2 will be implemented. However, at the subnational level, data sparsity is of great concern as even the methods that smooth over both age and period begin to suffer from large amounts of variability like the direct estimates. As cohort is less influenced by short-term fluctuations (unlike period) since it captures long-term (latent) trends, it is an ideal candidate to aid the production of smooth estimates and predictions. The smoother nature of cohort estimates when compared to a period estimate is seen when looking at the AP and AC national estimates. This highlights the importance of including cohort when one wishes to produce stable, smooth predictions.

As part of the analysis, we produce several metrics to assess whether the inclusion of cohort is appropriate. We assess the performance for both model fitting and model prediction. For the model fitting, the APC model was the best fitting model out of the APC, AP, and AC models

and for model prediction, whilst each of the three models are comparable to the direct estimates, the APC model was the closest on average. Therefore, by comparing against other models within the APC hierarchy and the well-known direct estimates, we have shown that the novel inclusion of cohort alongside age and period for smoothing aids in the production of smooth subnational estimates of U5MR.

The aim of this paper is to produce subnational estimates for U5MR that aid in policy and intervention decisions to achieve the SDG 3.2. The current model is tested on one of the better datasets the DHS has to offer at what is known as an Admin-1 subnational level. Since any interventions are implemented on the Admin-2 level, a finer subnational scale than the Admin-1 level, predictions at this level and for a wider range of datasets are required. In addition, the models currently used in the literature allow for multiple surveys for one country to be used. This is a very achievable extension as it would follow a well-defined approach (Wakefield et al., 2019).

A longer-term goal involves model selection and in particular, interaction term selection. Across the smoother methods for U5MR, a space-time interaction is often included but there has been little research conducted into other temporal-spatial and temporal-temporal interactions. The inclusion of additional temporal-temporal interactions may raise additional (lack-of) identification issues, but if the linear trends in the model are identifiable in the first place, these interactions can be included (Smith, 2018). The main work of this extension is the development of a thorough selection criteria to assess if the additional insight is worth the added complexity and computational cost to include the interaction.

To summarise, the novel inclusion of cohort alongside age and period when modelling U5MR is a suitable method for producing smooth subnational estimates. Immediate extensions include the use of multiple surveys for one country and application to an increased number of countries from the DHS and production of estimates on a finer subnational scale. Longer term goals include the development of a model selection procedure for additional interactions.

4.6 Chapter conclusions

In this chapter, we presented a novel application of our APC model to find subnational estimates of U5MR for Kenya using the 2014 KDHS survey data. The application included several extensions to the model seen in the previous chapters including, being fit to data in non-constant unequal intervals and the inclusion of terms for the urban/rural stratification, spatial location, and the space-time interaction. Cohort has not been included in previous models for U5MR due to added complications from the structural link and curvature identification problems. However, as the results show, cohort is an important inclusion in the model since it produces better fitting subnational estimates and predictions than when compared to models without cohort. We validated our model by comparing the results against those from the literature at the national level and during a LOOCV process. For both estimation and prediction, we defined several comparison metrics, and the APC model consistently scored the best. We showed the application of an APC model to U5MR is important as including cohort led to better fitting subnational estimates and predictions, a priority for the UN to achieve the SDG 3.2

Whilst the results of this initial application are encouraging, to have the model in a place where it can be used for policy implementation, additional work must be conducted. As there are multiple surveys from each LMIC country considered by the DHS, extending the current model to include multiple surveys will help reduce the data sparsity and give more informative inference. The DHS collect surveys from multiple countries and the 2014 KDHS is considered one of their better datasets. To ensure the model produces adequate estimates for a range of different dataset, we need to consider other countries. We include an interaction between spatial location and period to be in line with the literature, but a number of other interactions have not been included which can help further explain the uncertainty. Finally, we produce estimates at the Admin-1 level instead of the Admin-2 level where interventions occur. Whilst the initial results indicate the APC model is an appropriate novel extension, each of these additional extensions will further develop our model increasing its suitability for policy makers.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Overall conclusions

In this thesis, we have concerned ourselves with several different identification problems that arise when including the three temporal effects age, period, and cohort in one statistical model. Specifically, we considered the curvature identification problem that arises when an age-period-cohort (APC) model is fit to data that comes aggregated in unequal temporal intervals. The structural link identification problem of APC models is well known and has been considered on numerous occasions (Fienberg and Mason, 1979; Osmond and Gardner, 1982; Clayton and Schifflers, 1987). The most common solutions to the structural link identification problem involve reparameterising the model in terms of identifiable quantities, such as, temporal curvatures that are orthogonal to their respective temporal linear trends (Holford, 1983). The curvature identification problem has been considered far less in comparison (Holford, 2006). During this thesis, we set out to define a model that appropriately deals with the curvature identification problem and that can be used practically for public health research.

In Chapter 2, we showed that when considering data in unequal intervals, the previously identifiable terms (such as the temporal curvatures) are now unidentifiable. In addition, using penalised smoothing splines, we showed how a penalty is vital to address the curvature identification problem that arises when fitting APC models to data in unequal intervals. With theoretical illustrations, we showed how to reparameterise the penalty in line with the curvature terms and that a penalised smoothing spline incurs a larger penalty in the presence of the curvature identification problem than when not in its presence. Using empirical results, we showed how the un-penalised smoothing spline models ability to alleviate the curvature identification problem is sensitive to the spline specification, unlike with the penalised smoothing spline model concluding the neces-

sity of the penalty. The results of Chapter 2 are currently undergoing reviewer comment changes and a preprint is available at Gascoigne and Smith (2021).

In Chapter 3, we considered the correspondence between penalised smoothing splines and random walk priors. We offered a practical review of the correspondence by highlighting the key theoretical relations and using empirical results. Random walk priors have been used as a solution to the cyclic pattern in the estimates when fitting APC models to data aggregated in unequal intervals without the direct consideration of the curvature identification problem. We detail how to implement random walk of order two (RW2) priors on the curvature terms and showed how the correspondence between penalised smoothing splines and random walk of order two priors still holds in the context of APC modelling. We conclude the random walk of order two model is implementing its own version of a penalty like that of a penalised smoothing spline, and it is the penalisation that makes the RW2 prior model suitable to alleviate the curvature identification problem. Chapter 3 built upon the work of Chapter 2 as we now have two clear and well justified methods for fitting APC models to data that is aggregated in unequal intervals. The results for this chapter are drafted for publication.

In Chapter 4, we have an application to model under-five mortality rates (U5MR) to produce subnational estimates for low-and-middle income countries using data from the Demographic and Health Surveys (USAID, 2019). Alongside the random walk prior model, we included both spatial and spatio-temporal covariates in the model to utilise the geographical relationship shared between locations that are closer to one another to produced better fitting estimates of U5MR. This novel application of an APC model is required since current methods in the literature do not include cohort (but include age and period) as this induces both the structural link and curvature identification problems. The inclusion of cohort in models for U5MR offered better fitting models and predictions when compared to models that did not include cohort. This could be due to cohort offering a stabilising effect to the predictions since it is a measure of long-term, latent trends unlike period which is influenced by short term 'shocks'. The stabilising nature of cohort is vital to produce subnational estimates where data is sparse as this often causes high variation in the estimates. The results of this chapter are a prepared manuscript being reviewed by the co-authors.

5.2 Future work

When considering the correspondence between penalised smoothing splines and smoothing priors, we used random walk priors as these have commonly been used in APC modelling. Alternative choices in priors can also be used for APC models, for example Gaussian Process (GP) priors (Chernyavskiy et al., 2018). The correspondence between penalised smoothing splines and GP priors has been discussed with a practical view (Miller et al., 2020) but has not been considered

in relation to APC models. In Chapter 3, we confirmed the correspondence was still valid in the presence of the structural link and curvature identification problems for random walk (RW) priors. Including an exploration of GP priors for APC models is not as simple as the RW prior extension as it becomes unclear how to orthogonalize a GP prior with respect to a linear trend. But, by considering the work of both Chernyavskiy et al. (2018) and Miller et al. (2020), this provides a starting point for any exploration.

The structural link identification problem was resolved by reparameterising each temporal term into a linear trend and a set of orthogonal curvatures (Holford, 1983). There are other reparameterisations that are similar in the sense that they are based off identifiable terms equivalent to the curvature, i.e., double differences (Kuang et al., 2008) or curvatures with reference terms (Carstensen, 2007). Due to the equivalence between the identifiable terms in each reparameterisation, it is not unreasonable to assume the main result, the necessity of a penalty, will hold for the other two reparameterisations. In addition, there are other reparameterisations that can be explored in a similar manner. For example, the curvature can be split into identifiable quadratic terms (accelerations) and higher-order terms (Rosenberg, 2019). This allows a parameteric interpretation of the quadratic term, i.e., how fast the temporal trends are changing. Further work to explore the curvature identification problem and the use of a penalty to resolve it in the equivalent and new reparameterisations is an immediate development from this thesis.

We have considered data in equal interval, constant unequal interval, and non-constant unequal interval. One data format we have not considered is temporally misaligned data. It is common for providers of health and demographic data to release their data in either short (1×1), medium (5×5) and long (10×10) intervals; in addition, even if all the datasets are of the medium length, they may not have the same grouping. When combining multiple datasets of different length or if the same length but not the same groups, there is a misalignment issue. In spatial modelling this can be resolved by assuming a continuous spatial surface and evaluating the model using a stochastic partial differential equation approach (Wilson and Wakefield, 2020). Similarly, to spatial misalignment, continuous temporal functions may be the answer to temporal misalignment; however, understanding what is and is not identifiable and any additional issues that may occur for both equally and unequally temporally misaligned data is an important future development.

It is common for APC models to be chosen using the model hierarchy shown in Figure 5-1 or an equivalent starting from either period or cohort (Clayton and Schifflers, 1987). Whilst this hierarchy allows for the comparison between models that capture the three temporal trends, it does not consider any of the following temporal interaction terms, $f_{AP}(a,p)$, $f_{AC}(a,c)$, $f_{PC}(p,c)$ and $f_{APC}(a,p,c)$. The structural link identification issue arises when including a univariate term for each of age, period, and cohort in one model. This could be avoided by, for example, dropping the explicit cohort term for an interaction such as $f_{AP}(a,p)$, which captures trends equivalent to cohort without inducing the structural link (Luo and Hodges, 2020). A thorough model selection

criterion that considers all the temporal interactions would be a beneficial addition to the field as it would give practitioners a scheme to follow that would provide the most parsimonious model for the given scenario.

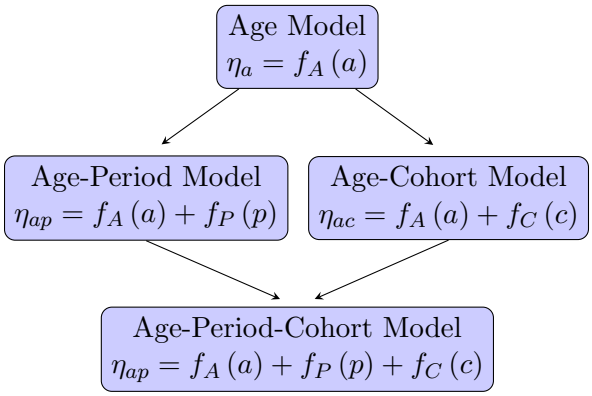


Figure 5-1: Classical APC model hierarchy.

For the spatio-temporal APC model to be fully included in specialist software for finding estimates of U5MR using survey data, there are three additional extensions that need to be considered: including more than one survey for a given country, testing on other countries, and finding estimates at a finer subnational level. Given most interventions occur at an Admin-2 level and we found estimates at an Admin-1 level, it is important to validate if the spatio-temporal APC model can produce estimates at the Admin-2 level where data sparsity is far more pronounced. Including the result of multiple surveys for the same country and alternative/additional interaction terms can help reduce the impact caused by data sparsity by explaining more uncertainty. In addition, the 2014 Kenyan DHS is considered one of the DHSs richer datasets. It is important to validate the spatio-temporal APC model performs as well as (if not better) than the current methods for all the DHS datasets, regardless of the quality of the data.

In this thesis, we showed how the novel inclusion of a penalty on the estimates of temporal curvature is critical to ensuring the curvature identification problem is alleviated. Due to this, we have shown both penalised smoothing splines and random walk of order two prior imposed on the curvature terms can define APC models that are appropriate for data aggregated in unequal intervals. Finally, we demonstrated the functionality of our model by using it in a novel application to find subnational estimates of U5MR.

BIBLIOGRAPHY

A. Baburin, R. Reile, T. Veideman, and M. Leinsalu. Age, period and cohort effects on alcohol consumption in Estonia, 1996–2018. *Alcohol and Alcoholism*, 56(4):451–459, 2021.

H. Bakka, H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018.

C. Berzuini and D. Clayton. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, 13(8):823–838, 1994.

C. Berzuini, D. Clayton, and L. Bernardinelli. Bayesian inference on the Lexis diagram. *Bulletin of the International Statistical Institute*, 55(1):149–165, 1993.

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, pages 3–41, 1995.

P. Boyle and C. Robertson. Statistical modelling of lung cancer and laryngeal cancer incidence in Scotland, 1960–1979. *American Journal of Epidemiology*, 125(4):731–744, 1987.

P. Boyle, N. E. Day, and K. Magnus. Mathematical modelling of malignant melanoma trends in Norway, 1953–1978. *American Journal of Epidemiology*, 118(6):887–896, 1983.

J. Cao, E. S. Eshak, K. Liu, A. Arafa, H. A. Sheerah, and C. Yu. Age-period-cohort analysis of stroke mortality attributable to high systolic blood pressure in China and Japan. *Scientific Reports*, 11(1):1–10, 2021.

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

Y.-Y. Chen, C.-T. Yang, E. Pinkney, and P. S. Yip. The Age-Period-Cohort trends of suicide in Hong Kong and Taiwan, 1979-2018. *Journal of Affective Disorders*, 295:587–593, 2021.

P. Chernyavskiy, M. P. Little, and P. S. Rosenberg. Correlated poisson models for age-period-cohort analysis. *Statistics in medicine*, 37(3):405–424, 2018.

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

L.-C. Chien, Y.-J. Wu, C. A. Hsiung, L.-H. Wang, and I.-S. Chang. Smoothed lexis diagrams with applications to lung and breast cancer trends in taiwan. *Journal of the American Statistical Association*, 110(511):1000–1012, 2015.

S. J. Clark, K. Kahn, B. Houle, A. Arteche, M. A. Collinson, S. M. Tollman, and A. Stein. Young children's probability of dying before and after their mother's death: A rural South African population-based surveillance study. *PLoS Medicine*, 10(3):e1001409, 2013.

D. Clayton and E. Schifflers. Models for temporal variation in cancer rates. II: Age–period–cohort models. *Statistics in Medicine*, 6(4):469–481, 1987.

T. J. Cole, C. Power, and G. E. Moore. Intergenerational obesity involves both the father and the mother. *The American Journal of Clinical Nutrition*, 87(5):1535–1536, 2008.

L. E. Eberly and B. P. Carlin. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, 19(17-18):2279–2294, 2000.

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

Z. Fannon, C. Monden, and B. Nielsen. 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)*, 184(3):842–867, 2021.

R. E. Fay and R. A. Herriot. Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

S. E. Fienberg and W. M. Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 10:1–67, 1979.

K. M. Flegal, M. D. Carroll, C. L. Ogden, and C. L. Johnson. Prevalence and trends in obesity among US adults, 1999-2000. *The Journal of the American Medical Association*, 288(14): 1723–1727, 2002.

C. Gascoigne and T. Smith. Using smoothing splines to resolve the curvature identifiability problem in age-period-cohort models with unequal intervals. *arXiv preprint arXiv:2112.08299*, 2021.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Third edition, 2013.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Routledge, 1990.

M. J. Haviland, A. Rowhani-Rahbar, and F. P. Rivara. Age, period and cohort effects in firearm homicide and suicide in the USA, 1983–2017. *Injury prevention*, 27(4):344–348, 2021.

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

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

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

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

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

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Kenya National Bureau of Statistics and Ministry of Health [Kenya] and Kenya, National AIDS Control Council [Kenya] and Kenya Medical Research Institute and National Council For Population And Development [Kenya] and The DHS Program, ICF International. Kenya demographic and health survey 2014, 2015. URL http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

N.-H. Kim and I. Kawachi. Age period cohort analysis of chewing ability in Korea from 2007 to 2018. *Scientific Reports*, 11(1):1–7, 2021.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

L. Knorr-Held. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567, 2000.

L. Knorr-Held and E. Rainer. Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics*, 2(1):109–129, 2001.

K. Kolpashnikova. Ageing and dementia: age-period-cohort effects of policy intervention in England, 2006–2016. *BMC geriatrics*, 21(1):1–6, 2021.

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

D. Kuh and Y. B. Shlomo. *A life course approach to chronic disease epidemiology*. Oxford University Press, Second edition, 2004.

I. Kuzmickiene and R. Everatt. Trends and age-period-cohort analysis of upper aerodigestive tract and stomach cancer mortality in Lithuania, 1987–2016. *Public Health*, 196:62–68, 2021.

Z. Li, Y. Hsiao, J. Godwin, B. D. Martin, J. Wakefield, S. J. Clark, with support from the United Nations Inter-agency Group for Child Mortality Estimation, and its technical advisory group. Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PloS one*, 14(1):e0210645, 2019.

Z. R. Li, B. D. Martin, T. Q. Dong, G.-A. Fuglstad, J. Godwin, J. Paige, A. Riebler, S. Clark, and J. Wakefield. *Space-Time Smoothing of Demographic and Health Indicators using the R Package SUMMER*, 2020.

F. Lindgren and H. Rue. On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35(4):691–700, 2008.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

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

L. Luo and J. S. Hodges. The age-period-cohort-interaction model for describing and investigating inter-cohort deviations and intra-cohort life-course dynamics. *Sociological Methods & Research*, page 0049124119882451, 2020.

Y. C. MacNab. On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, 20(1):49–68, 2011.

G. Martínez-Alés, J. R. Pamplin, C. Rutherford, C. Gimbrone, S. Kandula, M. Olfson, M. S. Gould, J. Shaman, and K. M. Keyes. Age, period, and cohort effects on suicide death in the United States from 1999 to 2018: Moderation by sex, race, and firearm involvement. *Molecular Psychiatry*, 26(7):3374–3382, 2021.

K. O. Mason, W. M. Mason, H. H. Winsborough, and W. K. Poole. Some methodological issues in cohort analysis of archival data. *American Sociological Review*, pages 242–258, 1973.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37. CRC press, Second edition, 1989.

L. Mercer, J. Wakefield, C. Chen, and T. Lumley. A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, 8:69–85, 2014.

L. Mercer, J. Wakefield, A. Pantazis, A. Lutambi, H. Masanja, and S. J. Clark. Small area estimation of child mortality in the absence of vital registration. *The Annals of Applied Statistics*, 9(4):1889–1905, 2015.

D. L. Miller, R. Glennie, and A. E. Seaton. Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, 25(1):1–16, 2020.

A. H. Mokdad, E. S. Ford, B. A. Bowman, W. H. Dietz, F. Vinicor, V. S. Bales, and J. S. Marks. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *The Journal of the American Medical Association*, 289(1):76–79, 2003.

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.

C. L. Ogden, M. D. Carroll, L. R. Curtin, M. A. McDowell, C. J. Tabak, and K. M. Flegal. Prevalence of overweight and obesity in the United States, 1999-2004. *The Journal of the American Medical Association*, 295(13):1549–1555, 2006.

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

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

C. Osmond and M. Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1(3):245–259, 1982.

C. Osmond and M. Gardner. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. *American Journal of Epidemiology*, 129(1):31–35, 1989.

J. Paige, G.-A. Fuglstad, A. Riebler, and J. Wakefield. Design-and model-based approaches to small-area estimation in a low-and middle-income country context: comparisons and recommendations. *Journal of Survey Statistics and Methodology*, 10(1):50–80, 2022.

A. L. Papoila, A. Riebler, A. Amaral-Turkman, R. São-João, C. Ribeiro, C. Geraldes, and A. Miranda. Stomach cancer incidence in southern Portugal 1998–2006: A spatio-temporal analysis. *Biometrical Journal*, 56(3):403–415, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

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

A. Riebler, L. Held, and H. Rue. Estimation and extrapolation of time trends in registry data—borrowing strength from related populations. *The Annals of Applied Statistics*, pages 304–333, 2012a.

A. Riebler, L. Held, H. Rue, and M. Bopp. Gender-specific differences and the impact of family integration on time trends in age-stratified Swiss suicide rates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2):473–490, 2012b.

A. Riebler, S. H. Sørbye, D. Simpson, and H. Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165, 2016.

C. Robertson and P. Boyle. Age, period and cohort models: The use of individual records. *Statistics in Medicine*, 5(5):527–538, 1986.

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

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

H. Rue and L. Held. *Gaussian Markov random fields: Theory and applications*. Chapman and Hall/CRC, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

M. J. Rutherford, P. C. Lambert, and J. R. Thompson. Age–period–cohort modeling. *The Stata Journal*, 10(4):606–627, 2010.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1): 1–28, 2017.

T. Smith. A stratified age-period-cohort model for spatial heterogeneity in all-cause mortality. *arXiv preprint arXiv:1806.02748*, 2018.

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

S. H. Sørbye and H. Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

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

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

UN IGME. *Levels & Trends in Child Mortality: Report 2021*. United Nations Inter-Agency Group for Child Mortality Estimation, https://childmortality.org/wp-content/uploads/2021/12/UNICEF-2021-Child-Mortality-Report.pdf, 2021.

United Nations. *Sustainable Development Goals*. http://sustainabledevelopment.un.org/owg.html, 2019.

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

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

J. Wakefield. *Bayesian and Frequentist Regression Methods*, volume 23. Springer, 2013.

J. Wakefield, G.-A. Fuglstad, A. Riebler, J. Godwin, K. Wilson, and S. J. Clark. Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 28(9):2614–2634, 2019.

J. Wakefield, T. Okonek, and J. Pedersen. Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418, 2020.

Z. Wang, E. Guo, B. Yang, R. Xiao, F. Lu, L. You, and G. Chen. Trends and age-period-cohort effects on mortality of the three major gynecologic cancers in China from 1990 to 2019: Cervical, ovarian and uterine cancer. *Gynecologic oncology*, 163(2):358–363, 2021.

S. Watanabe and M. Opper. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12):3571–3594, 2010.

K. Wilson and J. Wakefield. Pointless spatial modeling. *Biostatistics*, 21(2):e17–e32, 2020.

S. N. Wood. *Generalized additive models: An introduction with R*. Chapman and Hall/CRC, Second edition, 2017.

X. Wu, J. Du, L. Li, W. Cao, and S. Sun. Bayesian age-period-cohort prediction of mortality of Type 2 diabetic kidney disease in China: A modeling study. *Frontiers in endocrinology*, 12, 2021.

Y. R. Yue, P. L. Speckman, and D. Sun. Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics*, 64(3):577–613, 2012.

APPENDIX A

ADDITIONAL RESULTS FOR CHAPTER TWO

Here we include further results to supplement the binomial results for equal intervals displayed in Chapter 2. A more in depth look at the individual simulations for the binomial distribution is presented. Furthermore, results from the simulations for data generated under the Gaussian and Poisson distributions are included.

To show the structural link identification problem lies in the data, rather than in the choice of model fit, an additional simulation for the binomial distribution is included. The additional simulation study is for the full reparameterised APC models fit to data generated where only two of the three temporal effects are influential. The reparameterised APC model fit will have the cohort linear trend dropped and the temporal terms that influence the data are age and period.

A.1 Additional material for equal interval simulation

A.1.1 Individual simulations plot for the binomial distribution

Figures A-1-A-3 present the estimated functions from each simulation for the FA, RSS and PSS models for binomial data generated with all three effects present. In both the estimated effects and curvature plots, the FA and RSS models have a greater variability than the PSS model across all three temporal trends. The variability is much larger in the FA and RSS models for the youngest and oldest cohorts than for the PSS model. The earlier and later cohorts are seen the least throughout the data, hence suffer from greater variability; larger variability will incur a larger penalty in the PSS model.

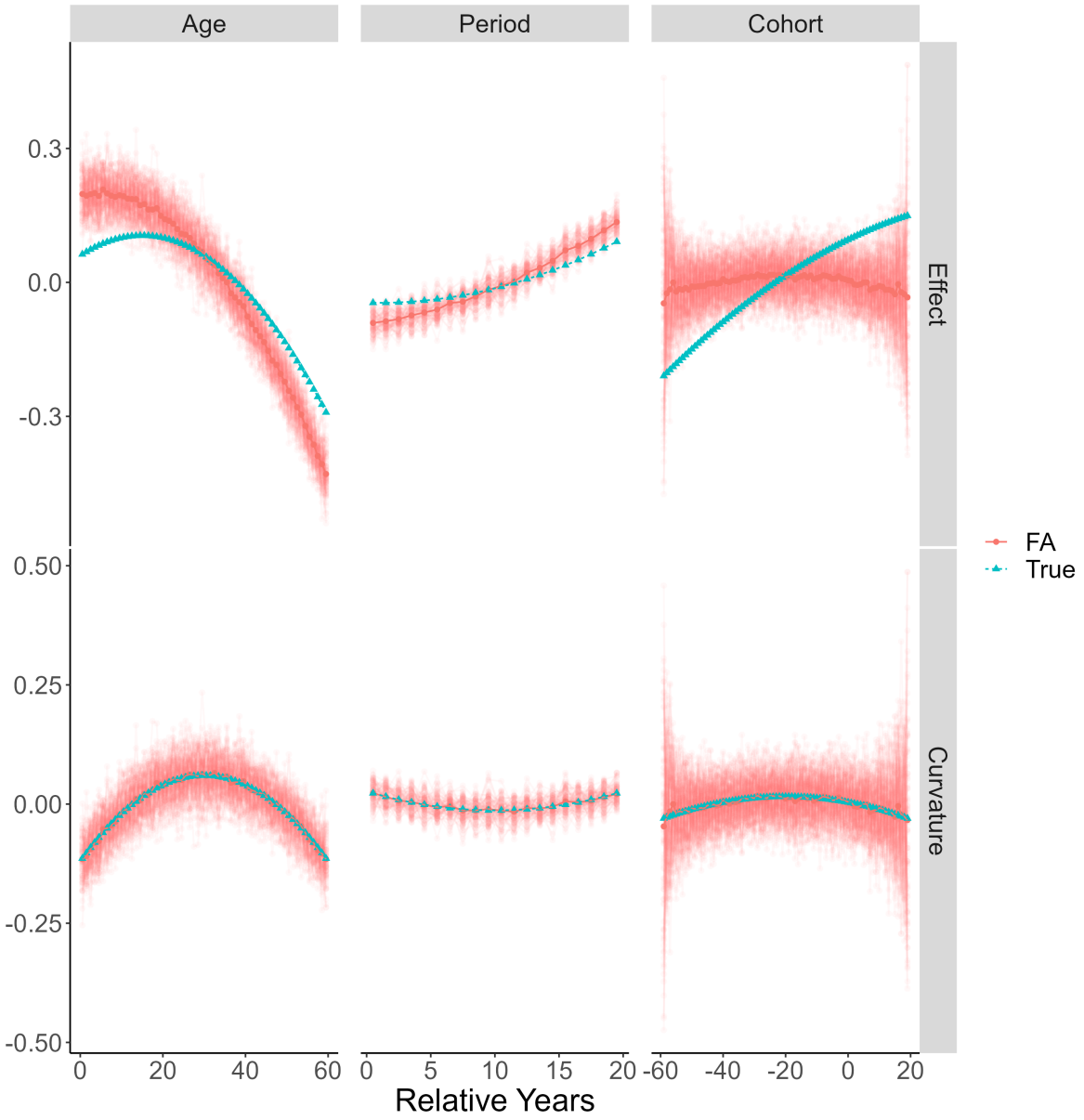


Figure A-1: Individual simulation plots for equal interval binomial data: factor model.

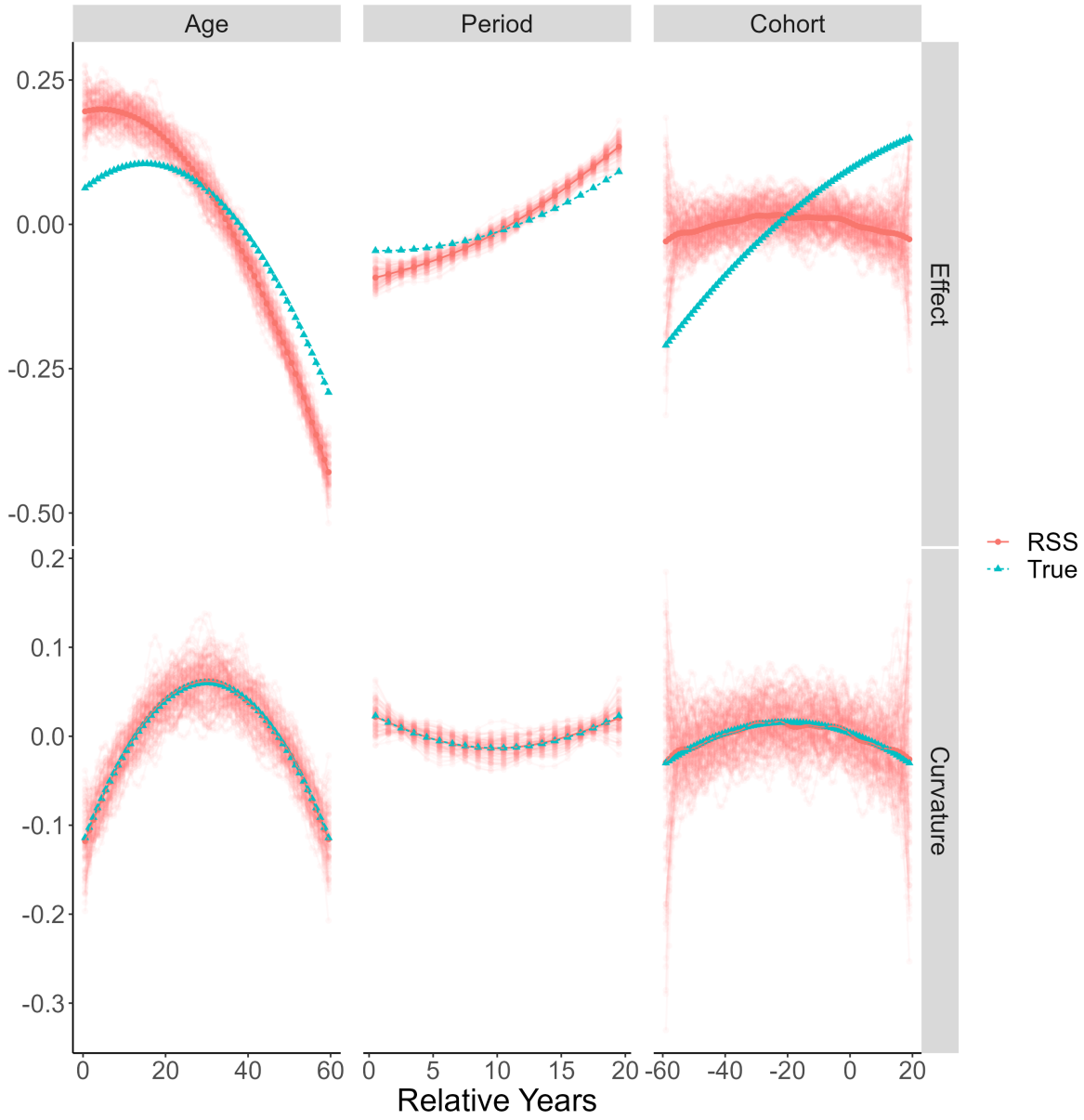


Figure A-2: Individual simulation plots for equal interval binomial data: regression smoothing spline model.

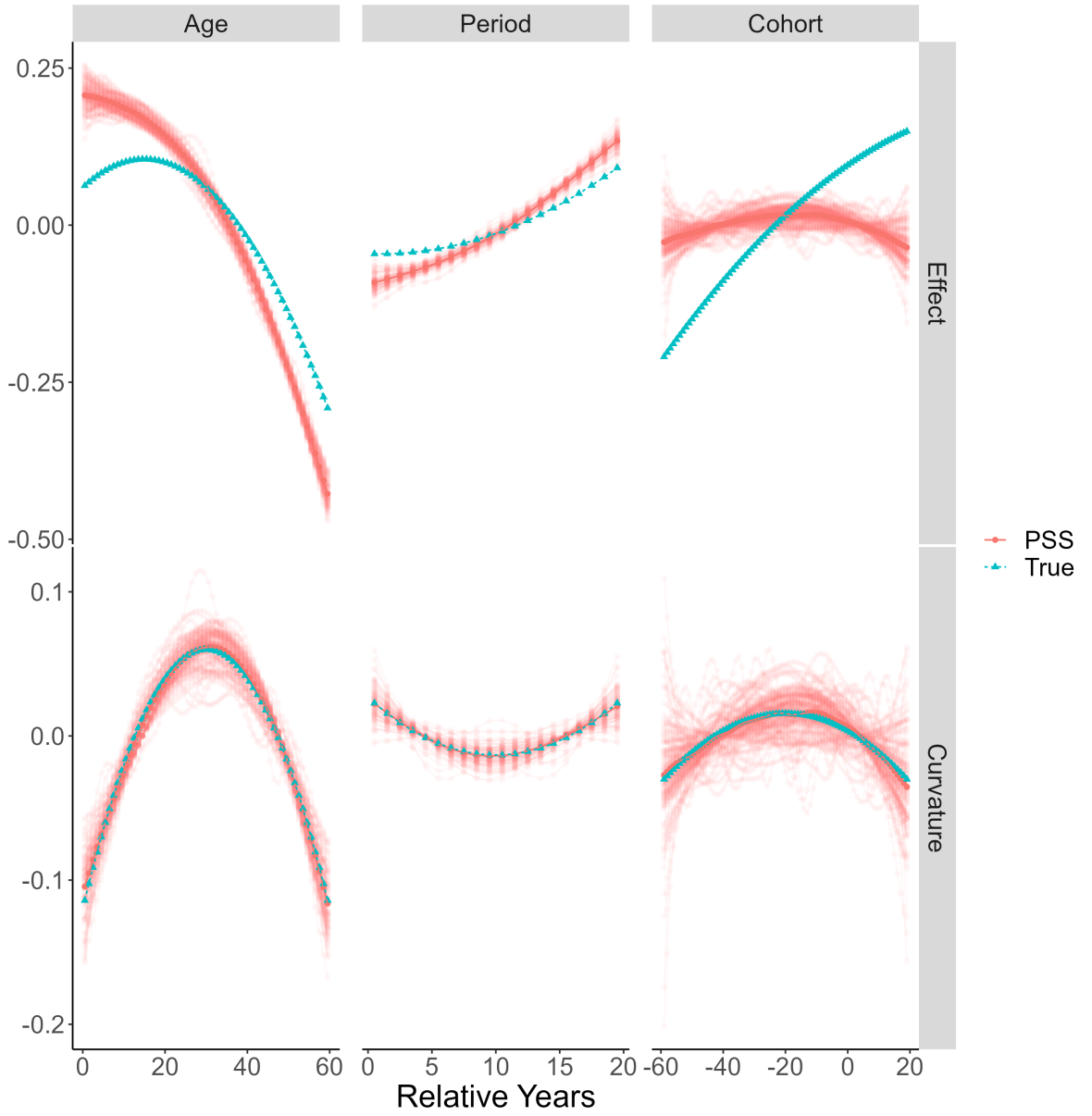


Figure A-3: Individual simulation plots for equal interval binomial data: penalised smoothing spline model.

A.1.2 Binomial simulation study for data generated with only two temporal effects present

Figure A-4 shows the simulation study results for equal interval, binomial data generated with only age and period effects present. Full APC models are fit to the data. In each model, cohort is the linear slope dropped in the reparameterisation.

The *ad-hoc* choice of what linear trend to drop is forcing that trend to be zero. When all three effects are present in the data generation, this is rarely the right choice as the true effect of the dropped trend is often not zero. In this simulation, we know cohort does not influence the data generated; therefore, the cohort linear trend is zero and the *ad-hoc* choice is correct. Th estimated effects correctly estimate the cohort linear trend and are shown to be the same as the true effects.

For the identifiable curvatures, the results for the FA and RSS models for the cohort have a relatively large bias and MSE in comparison to the PSS model. Each of the models is estimating a cohort curvature that is not present in the data; therefore, it is over-fitting the cohort curvature. The penalty term in the PSS model penalises the over-fitting hence why the cohort curvature bias and MSE is smaller in the PSS model than in the other two.

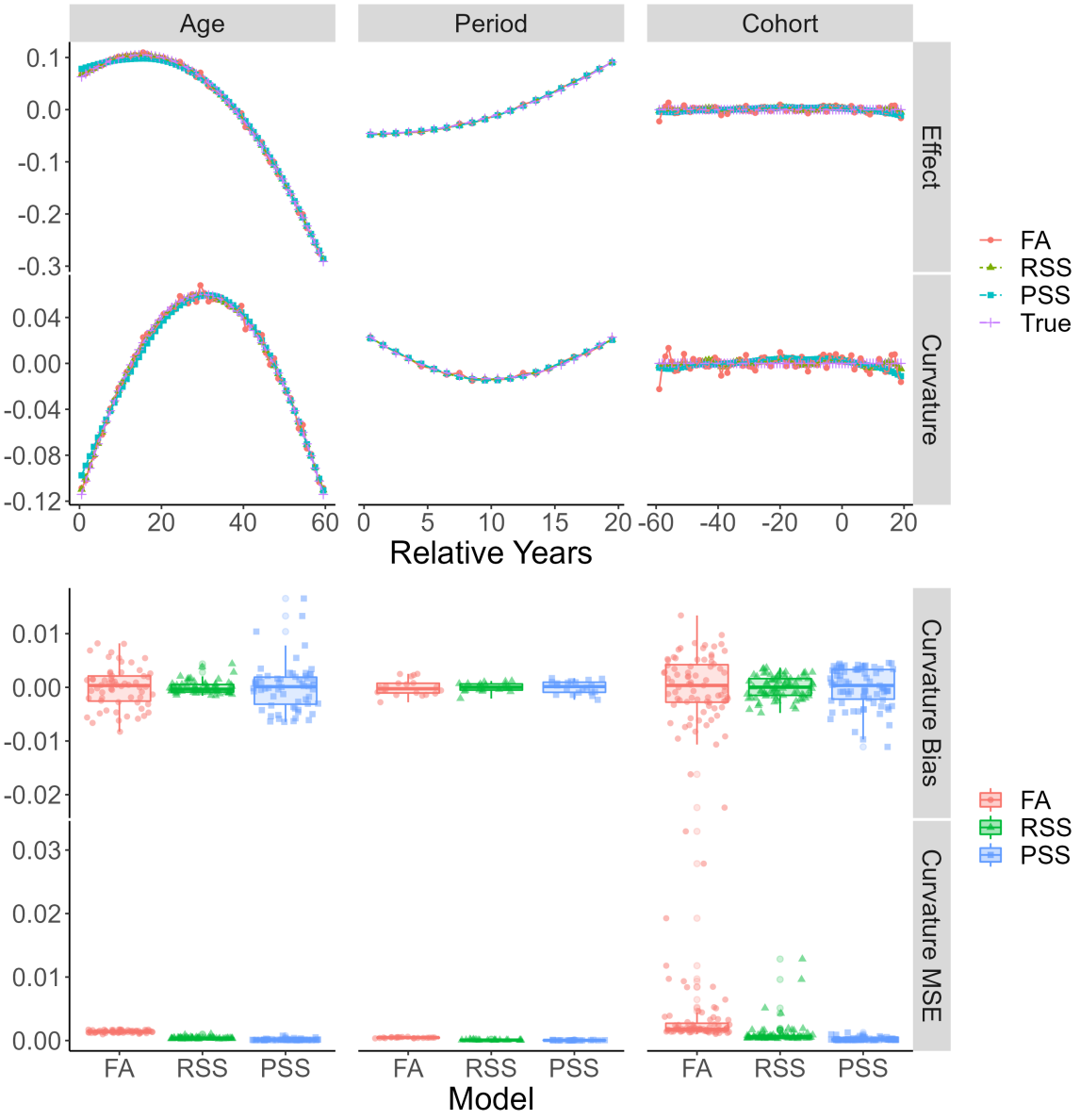


Figure A-4: Simulation study results for equal interval binomial data generated when only age and period effects are present.

A.1.3 Gaussian simulation studies

Figure A-5 shows the simulation study for equal interval, Gaussian data where all three effects are present. The results for Figure A-5 are like the results seen in the main body for the binomial distribution. The structural link identification issue is displayed by the model estimates for the temporal effects being different to the true effects. Furthermore, the curvatures are identifiable, and this is reflected in the model estimates of them matching the truth. The bias and MSE boxplots show the PSS model performs in line with the current literature.

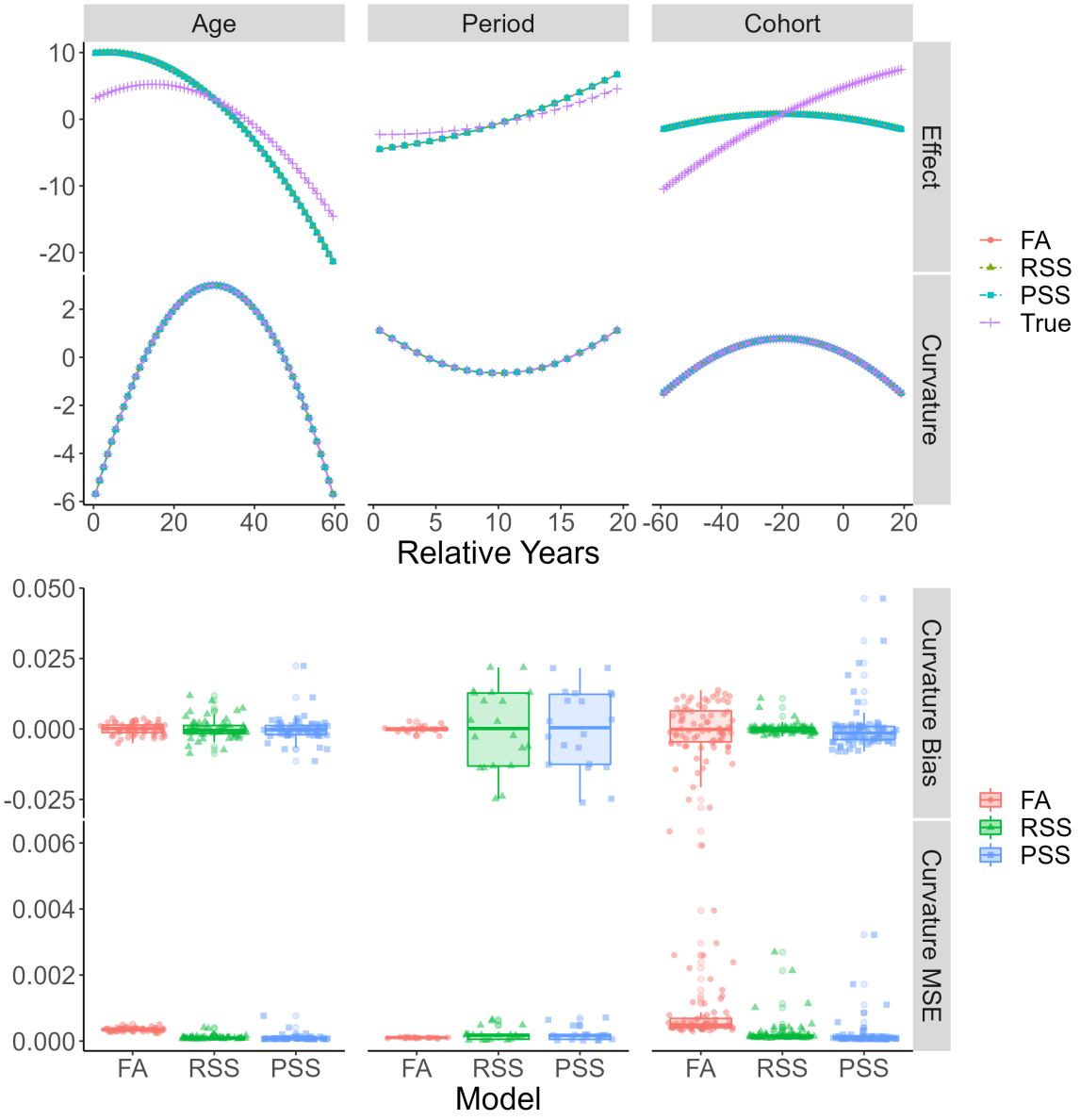


Figure A-5: Simulation study results for equal interval, $M = 1$, Gaussian data generated when all three temporal effects are present.

A.1.4 Poisson simulation studies

Figure A-6 shows the simulation study for equal interval, Poisson data where all three effects are present. The interpretation of the results from the Poisson simulation study is the same as those from the binomial and Gaussian distributions.

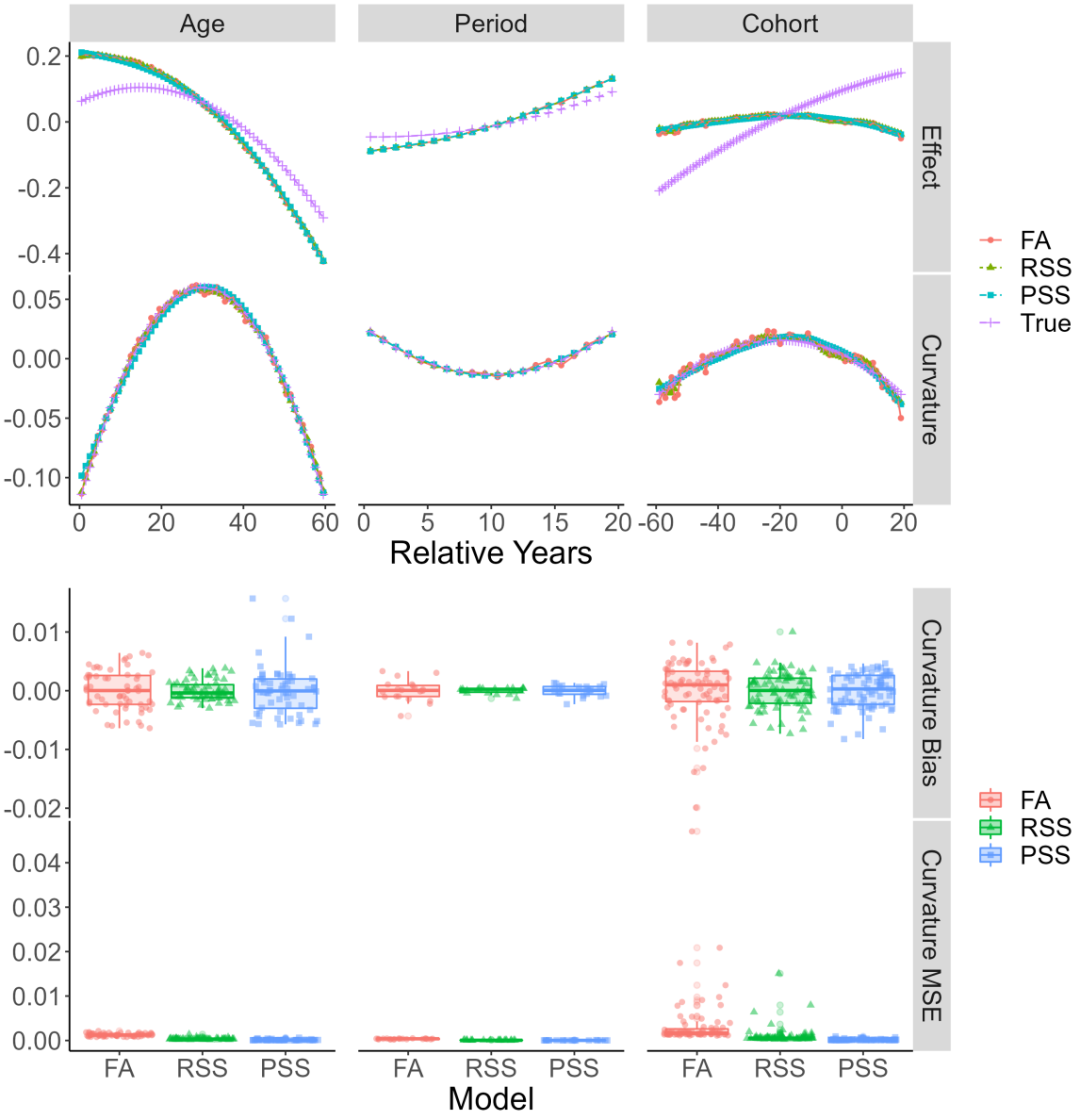


Figure A-6: Simulation study results for equal interval, $M = 1$, Poisson data generated when all three temporal effects are present.

A.2 Theoretical justification for the use of a penalty function

The reparameterised APC model penalty function from Section 2.2.4 with the transformed functions is

$$\lambda_A \int \tilde{f}''_{A_C}(a)^2\, da + \lambda_P \int \tilde{f}''_{P_C}(p)^2\, dp + \lambda_C \int \tilde{f}''_{C_C}(c)^2\, dc$$

where $\lambda_\star$ are the temporal smoothing parameters controlling the trade off between the estimates fit and smoothness. $\lambda_\star \to \infty$ and $\lambda_\star = 0$ lead to straight line and un-penalised estimates, respectfully. When the the integrated square of the second derivative is large, the smoothing parameter increases in order to apply additional cost to fitting the complicated function. Therefore, if the period and cohort curvature approximate $v_M \neq 0$, the larger integrated square of the second derivative, in comparison to when $v_M = 0$ is approximated, yields a greater cost to fit.

Due to the curvature identifiability problem, we propose for fixed λ_P and λ_C the estimates of the transformed period and cohort functions that approximate $v_M = 0$ have the smallest integrated squared second derivatives. This results in estimates for the transformed functions approximating $v_M = 0$ having the smallest penalty and

$$\begin{aligned}\lambda_P \int \tilde{f}''_{P_C}(p)^2\, dp + \lambda_C \int \tilde{f}''_{C_C}(c)^2\, dc &= \lambda_P \int \left[f''_{P_C}(p) + v''_{M_C}(p)\right]^2 dp + \lambda_C \int \left[f''_{C_C}(c) - v''_{M_C}(c)\right]^2 dc \\ &\geq \lambda_P \int f''_{P_C}(p)^2\, dp + \lambda_C \int f''_{C_C}(c)^2\, dc.\end{aligned}$$

When estimating APC trends for unequal data, the age curvature is still identifiable Eq.(2.5). Consequently, and for the purpose of explanation, we omit the age curvature penalty term from the above expression and in the below theoretical section. However, we would like to stress that the age curvature penalty is still present in the estimation process.

To demonstrate the penalty term for period and cohort curvature estimates is greater when $v_M \neq 0$ is approximated than for when $v_M = 0$ is approximated, consider the special case using a natural cubic spline with knots spaced M apart. Natural cubic splines enforce linearity beyond the boundary knots which reduces instability in these regions. This is a useful quality for the likes of cohort where the tails are based on few observations (Heuer, 1997). First, we consider period.

Both $\tilde{f}_{P_C}$ and f_{P_C} are defined by the same set of knots $\check{p}_1 < \cdots < \check{p}_J$ which are spaced M apart from one another. The natural cubic splines are continuous to the second derivatives (f'''_{P_C} is constant in and $\check{p}_j^+$ is a point in $[\check{p}_j, \check{p}_{j+1}]$) and are linear beyond the boundary knots ($f''_{P_C}(\check{p}_1) = f''_{P_C}(\check{p}_J) = 0$). Therefore, the penalty term of $\tilde{f}_{P_C}$ can be written as

$$\int_{\check{p}_1}^{\check{p}_J} \tilde{f}''_{P_C}(p)^2\, dp = \int_{\check{p}_1}^{\check{p}_J} f''_{P_C}(p)^2\, dp + 2\int_{\check{p}_1}^{\check{p}_J} f''_{P_C}(p)\, v''_{M_C}(p)\, dp + \int_{\check{p}_1}^{\check{p}_J} v''_{M_C}(p)^2\, dp$$

when integrated over the range of all knots.

Applying integration by parts to the cross term

$$\begin{aligned}
\int_{\check{p}_1}^{\check{p}_J} f''_{P_C}(p)\, v''_{M_C}(p)\, dp &= f''_{P_C}(p)\, v'_{M_C}(p)\big|_{\check{p}_1}^{\check{p}_J} - \int_{\check{p}_1}^{\check{p}_J} f'''_{P_C}(p)\, v'_{M_C}(p)\, dp \\
&= -\int_{\check{p}_1}^{\check{p}_J} f'''_{P_C}(p)\, v'_{M_C}(p)\, dp \\
&\quad \text{as } f''_{P_C}(\check{p}_1) = f''_{P_C}(\check{p}_J) = 0 \\
&= -\sum_{j=1}^{J-1} f'''_{P_C}\left(\check{p}_j^{+}\right) \int_{\check{p}_j}^{\check{p}_{j+1}} v'_{M_C}(p)\, dp \\
&\quad \text{as } f'''_{P_C} \text{ is constant and } \check{p}_j^{+} \text{ is a point in } [\check{p}_j, \check{p}_{j+1}] \\
&= -\sum_{j=1}^{J-1} \text{constant}\left[v_{M_C}(\check{p}_{j+1}) - v_{M_C}(\check{p}_j)\right] \\
&= 0
\end{aligned}$$

since $\check{p}_j$ and $\check{p}_{j+1}$ are M apart ($v_{M_C}(\check{p}_{j+1}) = v_{M_C}(\check{p}_j)$). We have shown that

$$\begin{aligned}
\int_{\check{p}_1}^{\check{p}_J} \tilde{f}''_{P_C}(p)^2\, dp &= \int_{\check{p}_1}^{\check{p}_J} f''_{P_C}(p)^2\, dp + \int_{\check{p}_1}^{\check{p}_J} v''_{M_C}(p)^2\, dp \\
&\geq \int_{\check{p}_1}^{\check{p}_J} f''_{P_C}(p)^2\, dp
\end{aligned}$$

as $\int_{\check{p}_1}^{\check{p}_J} v''_{M_C}(p)^2\, dp \geq 0$.

By applying the same reasoning as we did with period, we can similarly show that

$$\begin{aligned}
\int_{\check{c}_1}^{\check{c}_K} \tilde{f}''_{C_C}(c)^2\, dc &= \int_{\check{c}_1}^{\check{c}_K} f''_{C_C}(c)^2\, dc + \int_{\check{c}_1}^{\check{c}_K} v''_{M_C}(c)^2\, dc \\
&\geq \int_{\check{c}_1}^{\check{c}_K} f''_{C_C}(c)^2\, dc.
\end{aligned}$$

As a results, we have shown that when the curvature identifiability problem is present, the case when the period and cohort curvature functions is approximating $v_M \neq 0$, the integrated square of the second derivative of the period and cohort curvature estimates is greater than when the functions are approximating $v_M = 0$.

A.3 Additional material for unequal interval simulation

Here we include further results to supplement the binomial results for unequal intervals displayed in Chapter 2. As with the additional results for the equal intervals simulation we have: individual simulations for the binomial distribution; an additional simulation study for the binomial distribution when the data is generated with only two temporal effects; and results from the Gaussian and Poisson distributions for data generated with all three temporal effects.

A.3.1 Individual simulations plot for the binomial distribution

Figures A-7-A-9 present the estimated functions from each simulation for the FA, RSS and PSS models for unequal interval, binomial data generated with all three effects present.

Figure A-7 shows how the FA model is unsuitable for data that comes in unequal intervals. The large variability throughout each of the estimated curvature functions highlights how the previously identifiable terms are no longer identifiable. Furthermore, the age curvatures not suffering the same issues as the period and cohort curvatures shows how the additional identification from fitting a model to unequally aggregated data only affects the period and cohort terms.

The RSS model has a larger variation for its simulations than the PSS model as well as suffering from greater variability in the tails of each function, most notably for cohort. This shows how the penalty term being applied in the PSS model is working to reduce the additional identification in the curvature estimates for each individual simulation.

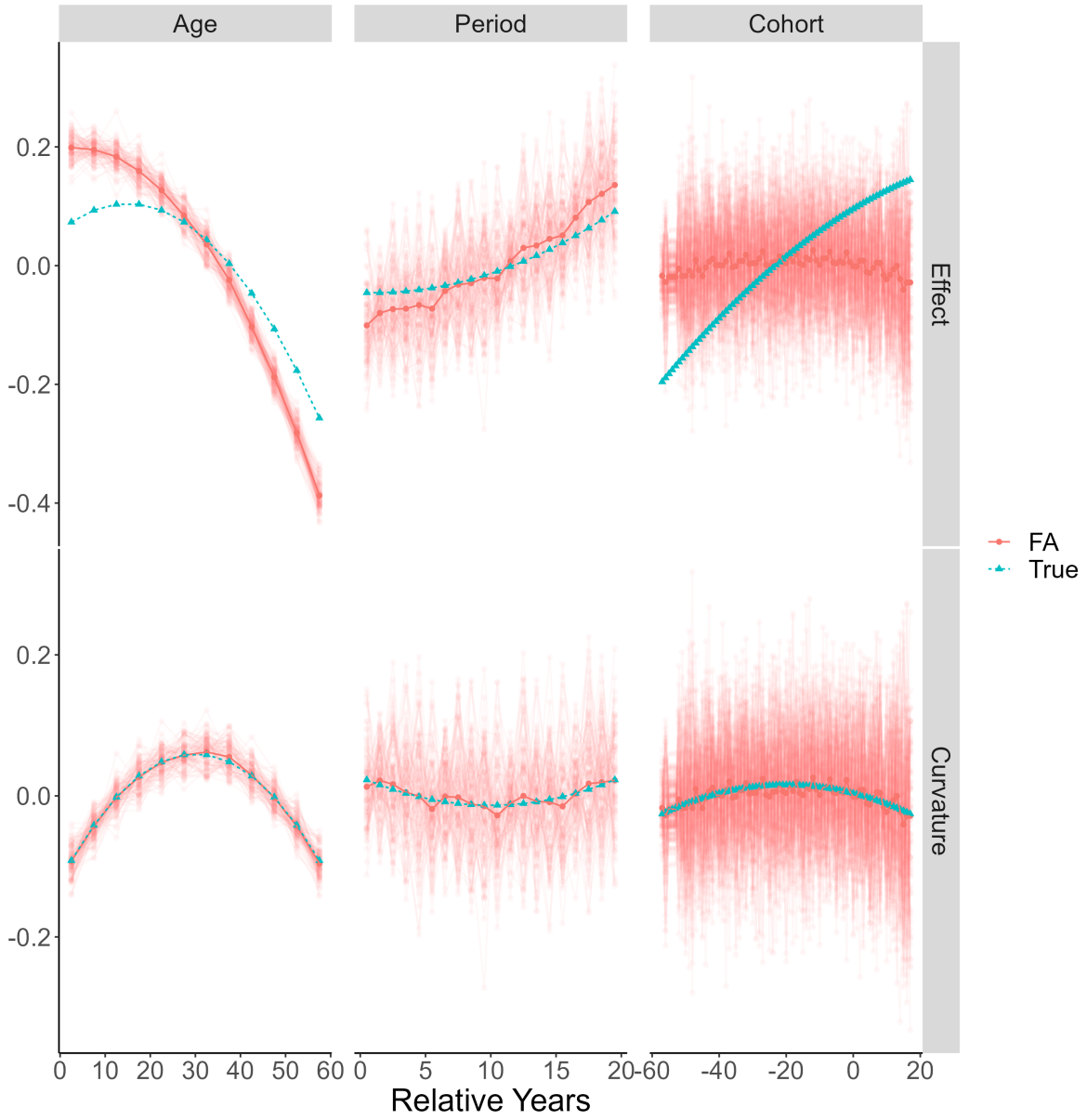


Figure A-7: Individual simulation plots for unequal interval binomial data: factor model.

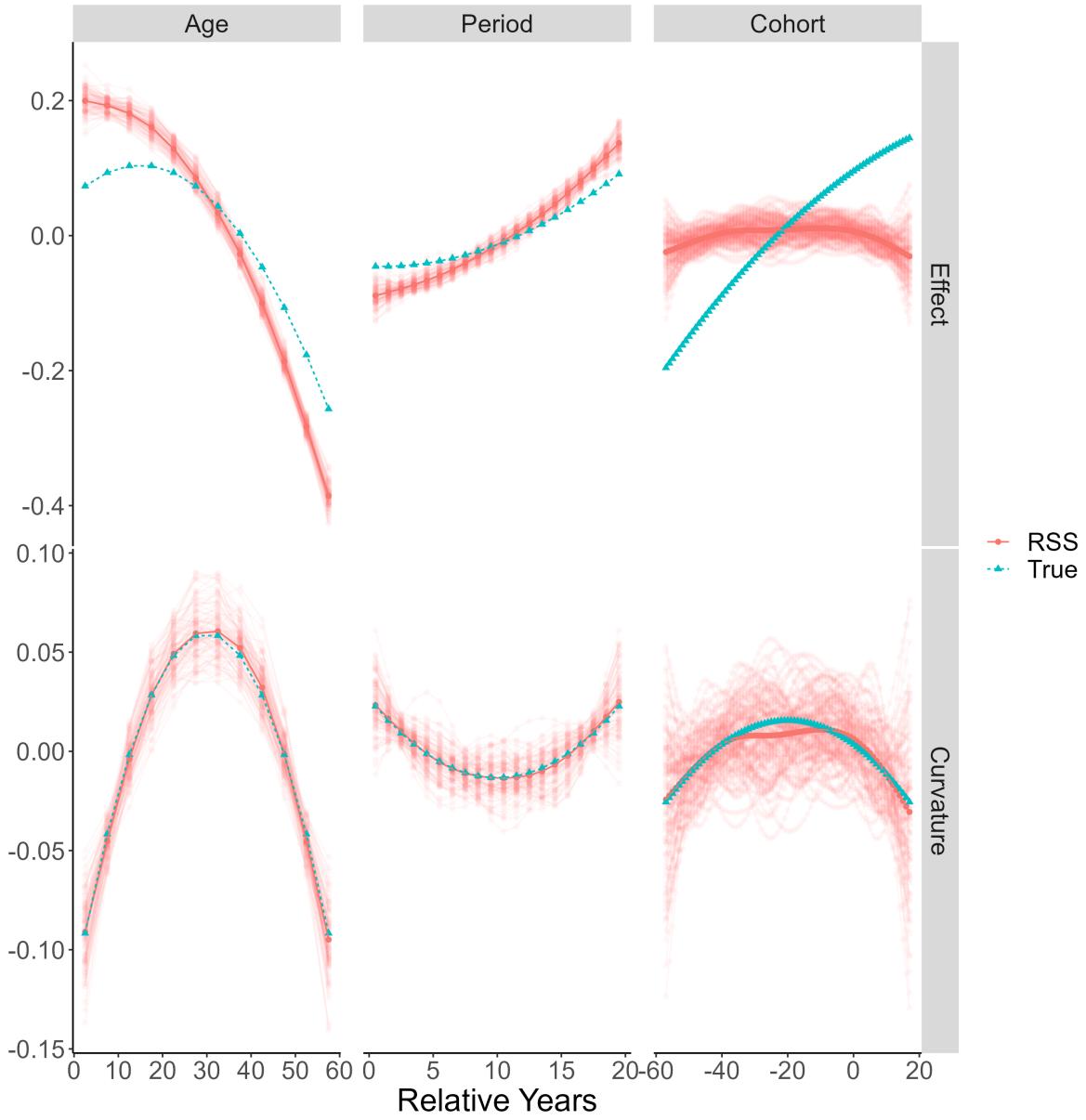


Figure A-8: Individual simulation plots for unequal interval binomial data: regression smoothing spline model.

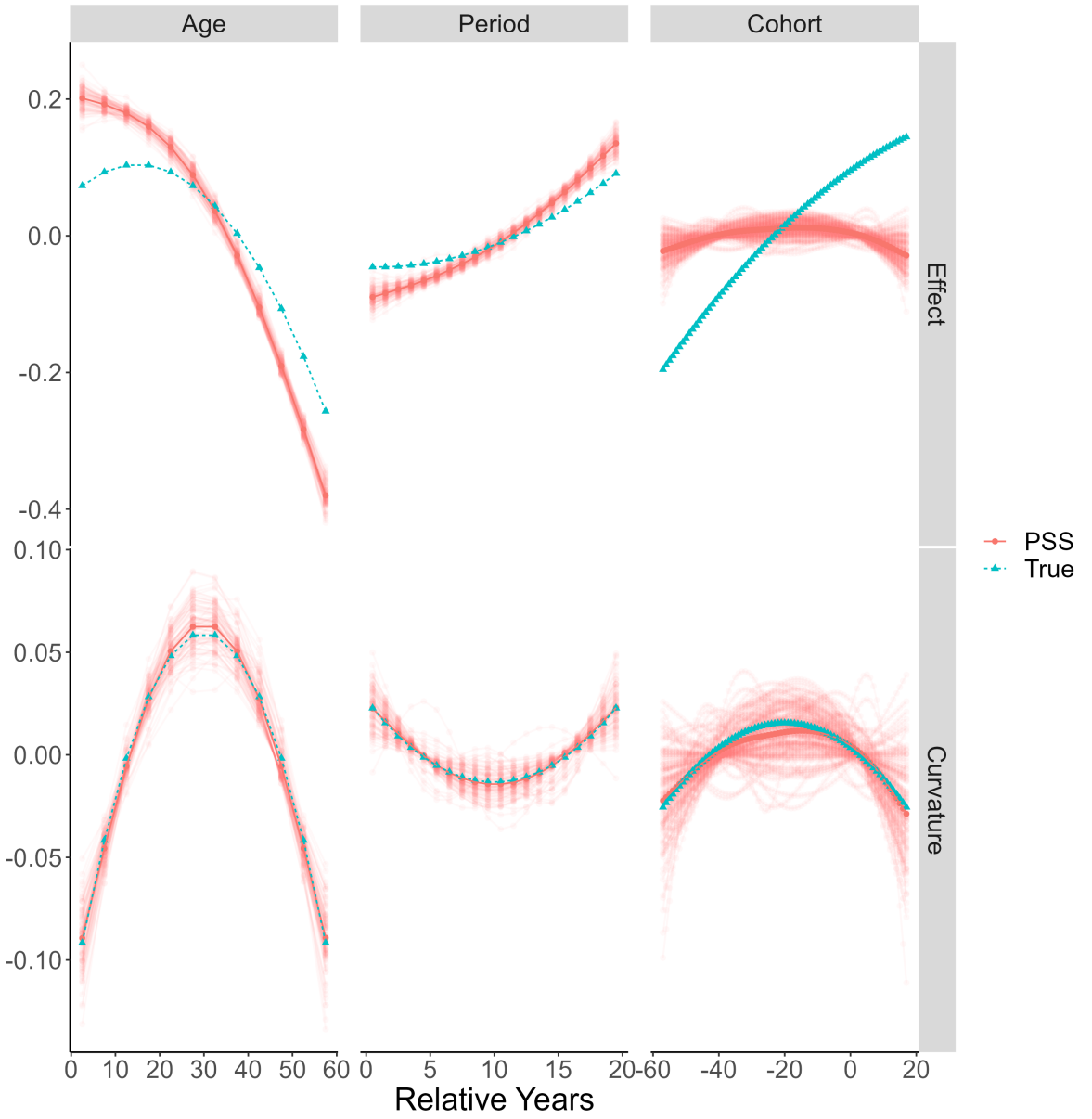


Figure A-9: Individual simulation plots for unequal interval binomial data: penalised smoothing spline model.

A.3.2 Binomial simulation study for data generated with only two temporal effects present

Figure A-10 presents the results from the simulation study where only age and period are influence the data generation. As with the equal intervals, the structural link identification problem is no longer present in the data hence why the estimated effects appear to be identifiable at first look.

Looking closer, the FA model estimates are exhibiting the periodic pattern in the period and cohort estimated effects due to the additional identification issues (Holford, 2006). The periodic pattern in the FA model becomes more pronounced in the curvature estimates. A smaller bias for the PSS period and cohort curvature box-plots in comparison to the RSS period and cohort curvature bias boxplots shows the PSS model is performing better than the RSS model. The additional identification problems in the estimated curvature functions are being alleviated by the PSS model; the penalty is resolving the additional identification issues in the estimates whereas the use of smoothing functions alone is not.

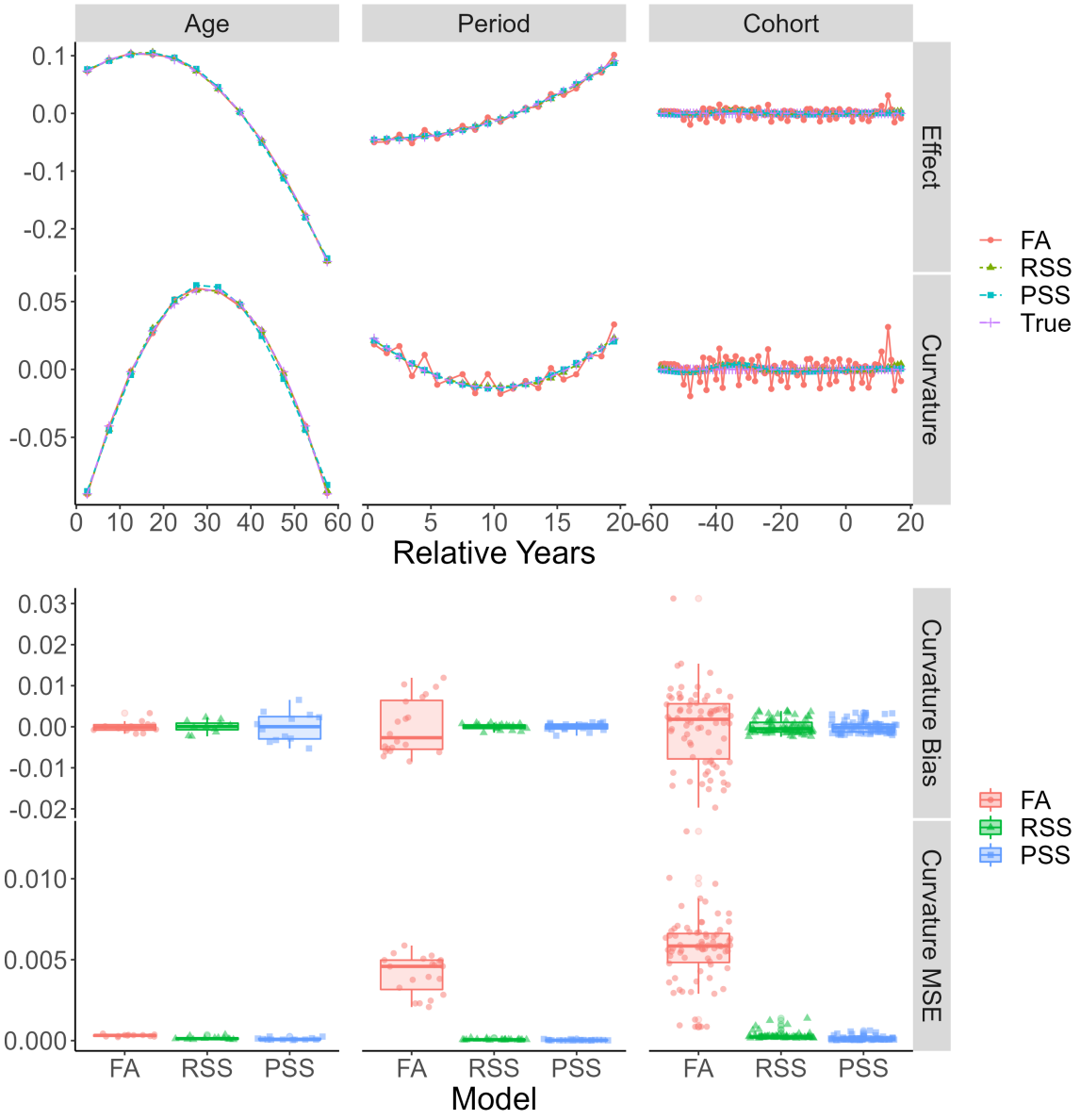


Figure A-10: Simulation study results for unequal interval binomial data generated when only age and period effects are present.

A.3.3 Gaussian simulation studies

Figure A-11 presents the simulation study for unequal interval, normal data generated with all three temporal effects present.

The results from Figure A-11 show the FA model displaying the cyclic pattern of the identification issues due to fitting an APC model to the unequally aggregated APC data. As with the binomial results, the RSS and PSS results are hard to distinguish between, but this is due to the choice of basis function used to approximate the true function. If the chosen basis function is not a good approximation to the true function, the estimates will not display the cyclic pattern of the added identification and the difference between the methods is hard to tell. When the basis is more informative (spans a larger space of the true function), the strengths of the PSS model become apparent; the PSS model gives the user confidence the results they have are not as influenced by the choice in basis as the RSS result. The penalisation of the estimates in the PSS model will always alleviate the added identification no matter the basis, this is not the case for the RSS model.

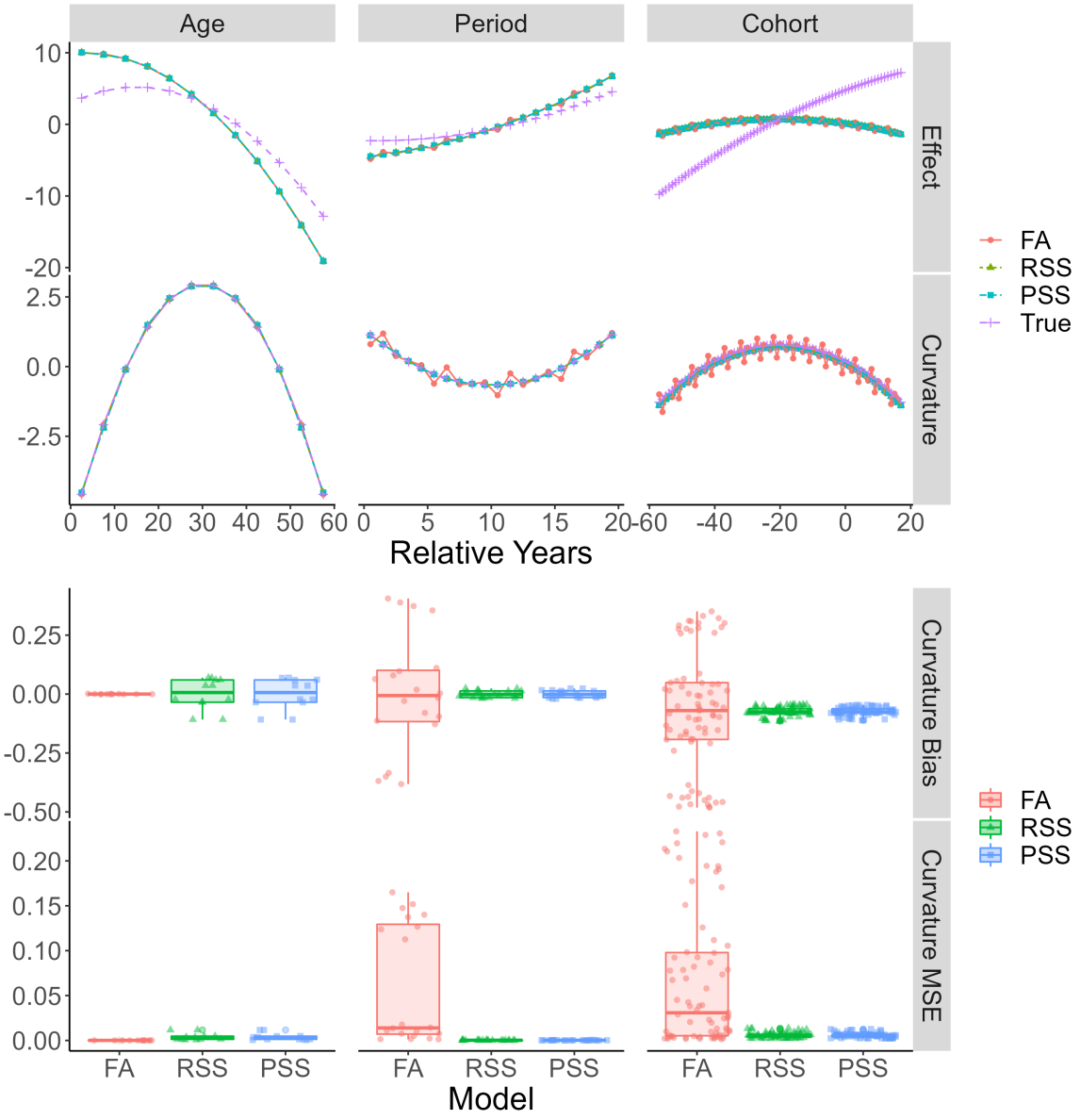


Figure A-11: Simulation study results for unequal interval, $M = 5$, Gaussian data generated when all three temporal effects are present.

A.3.4 Poisson simulation studies

Figure A-12 shows the simulation study for unequal interval, Poisson data with all three temporal effects present. The interpretation of the results from the Poisson simulation studies is the same as those from the binomial and Gaussian distributions.

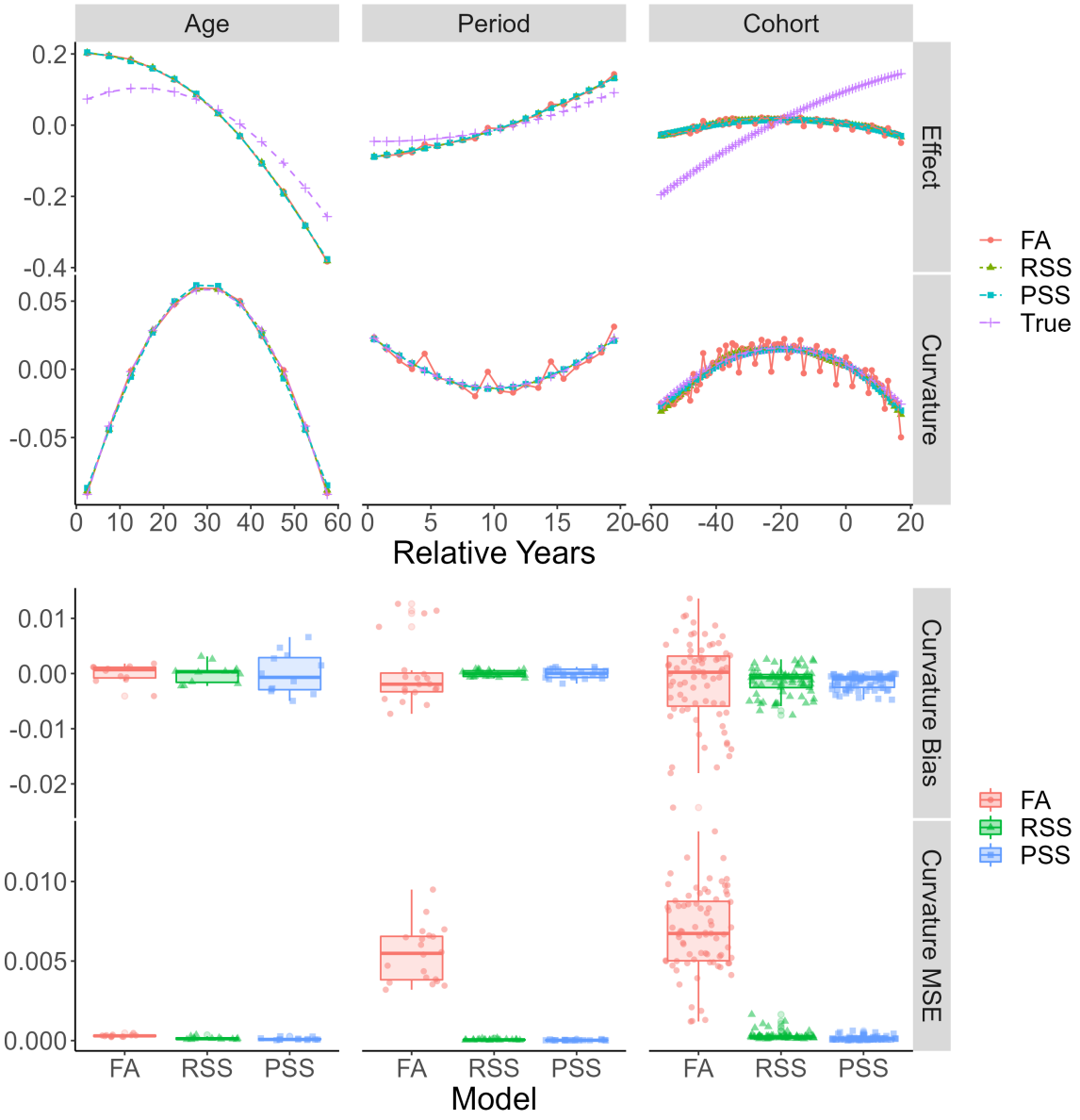


Figure A-12: Simulation study results for unequal interval, $M = 5$, Poisson data generated when all three temporal effects are present.

A.4 Application

Figures A-13-A-16 show the true and predicted heat maps of all-cause mortality for ages 0-99 and years 1925-2015 in the UK for the different data aggregation formats. The gradient change between dark blue to red shows the mortality rate increasing. Figure A-13 is the true rate and Figures A-14 - A-16 are the predicted rates for the 1×1, 5×1 and 5×5 data sets. Each model accurately predicts the overall trends with similar mortality rates across the whole map when compared to the true rate. The differences in pixel size for each of the predicted maps is due to how the data is formatted. As expected, collapsing 5×1 data into 5×5 data results in a loss of information and this can be seen by the increase in pixel size along the x-axis when comparing between Figure A-15 and Figure A-16.

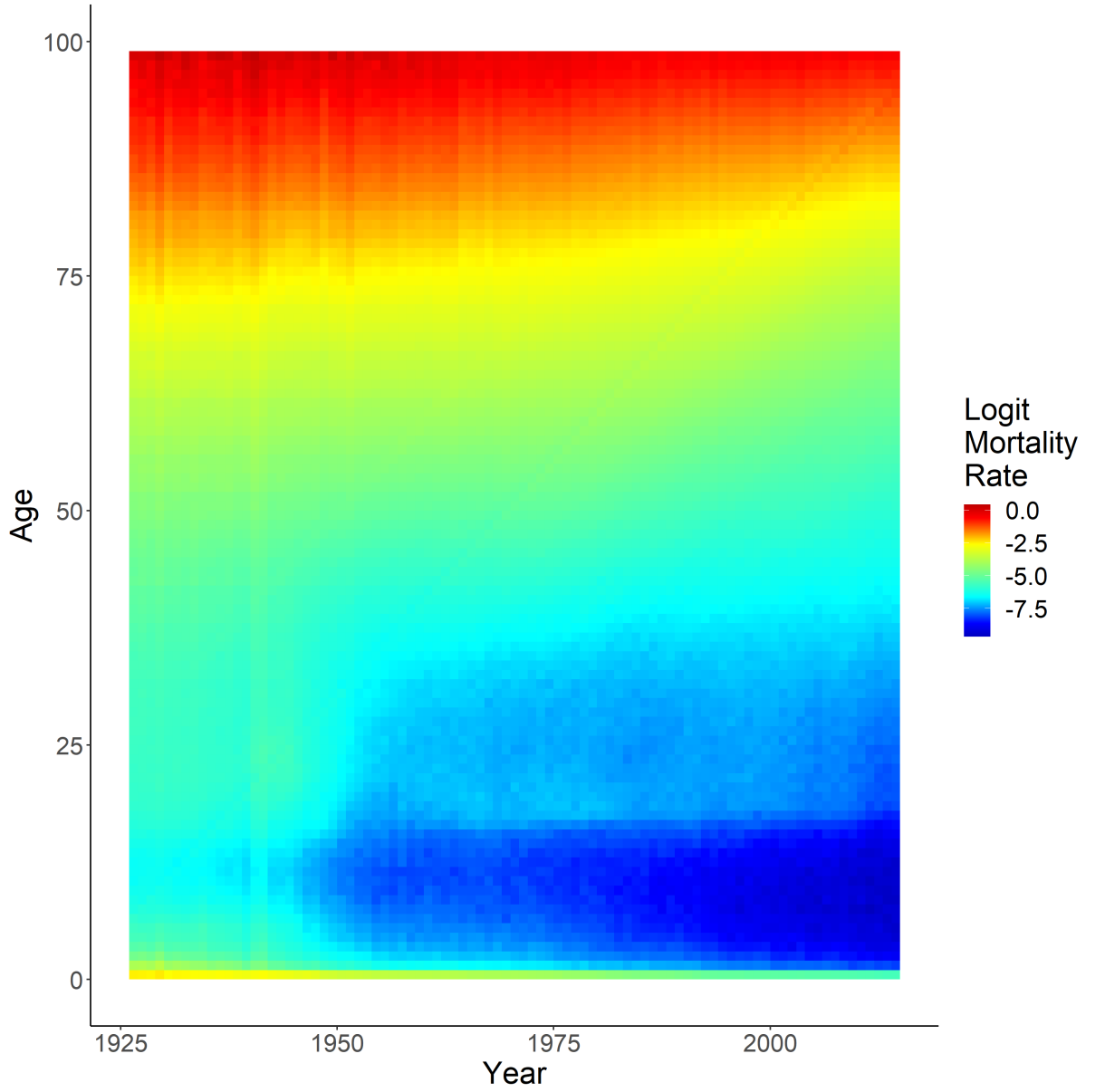


Figure A-13: True all-cause mortality for years 1926-2015 and ages 0-99 in the United Kingdom for data aggregated in 1×1.

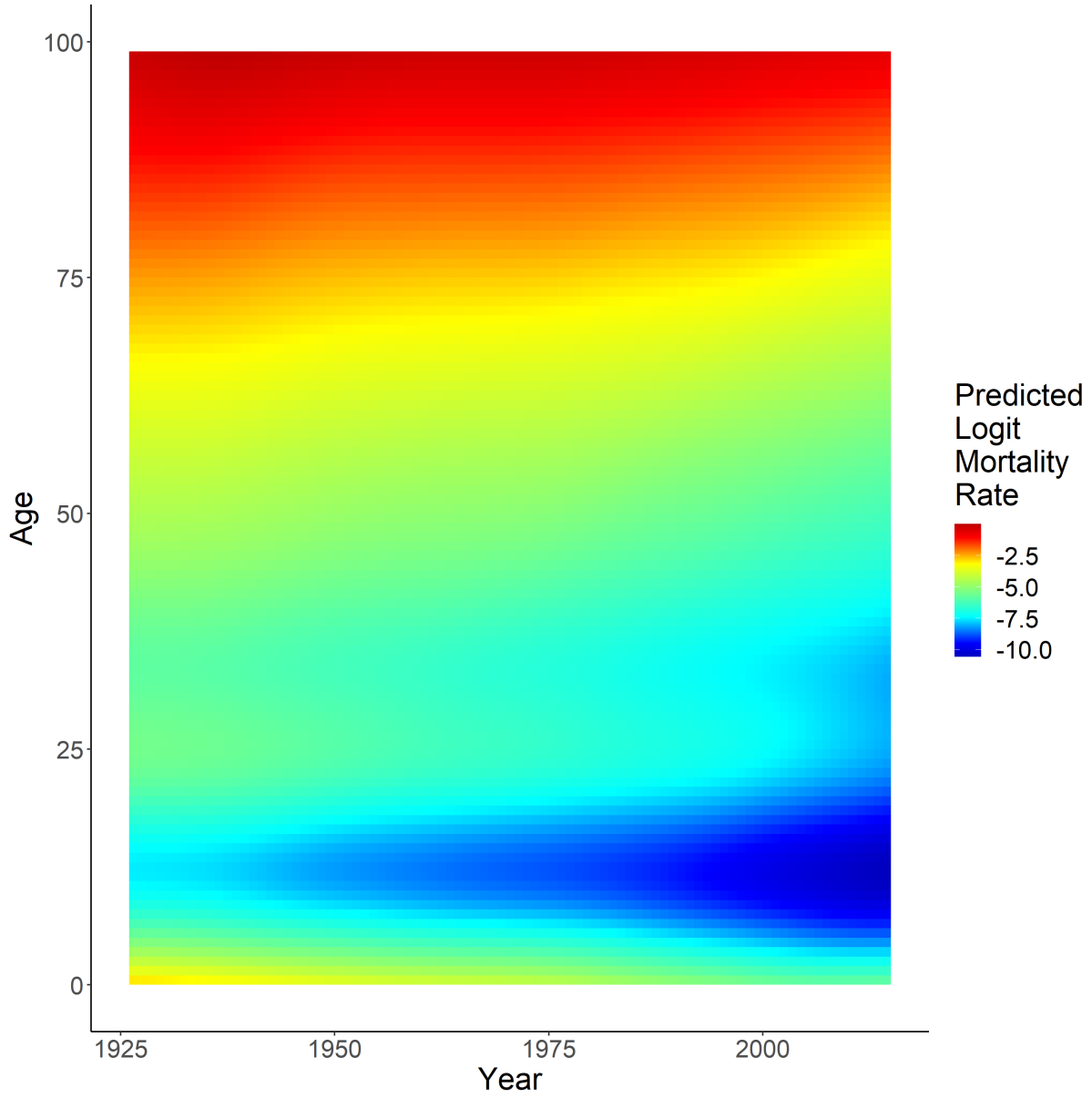


Figure A-14: Predicted all-cause mortality for years 1926-2015 and ages 0-99 in the United Kingdom for data aggregated in 1×1.

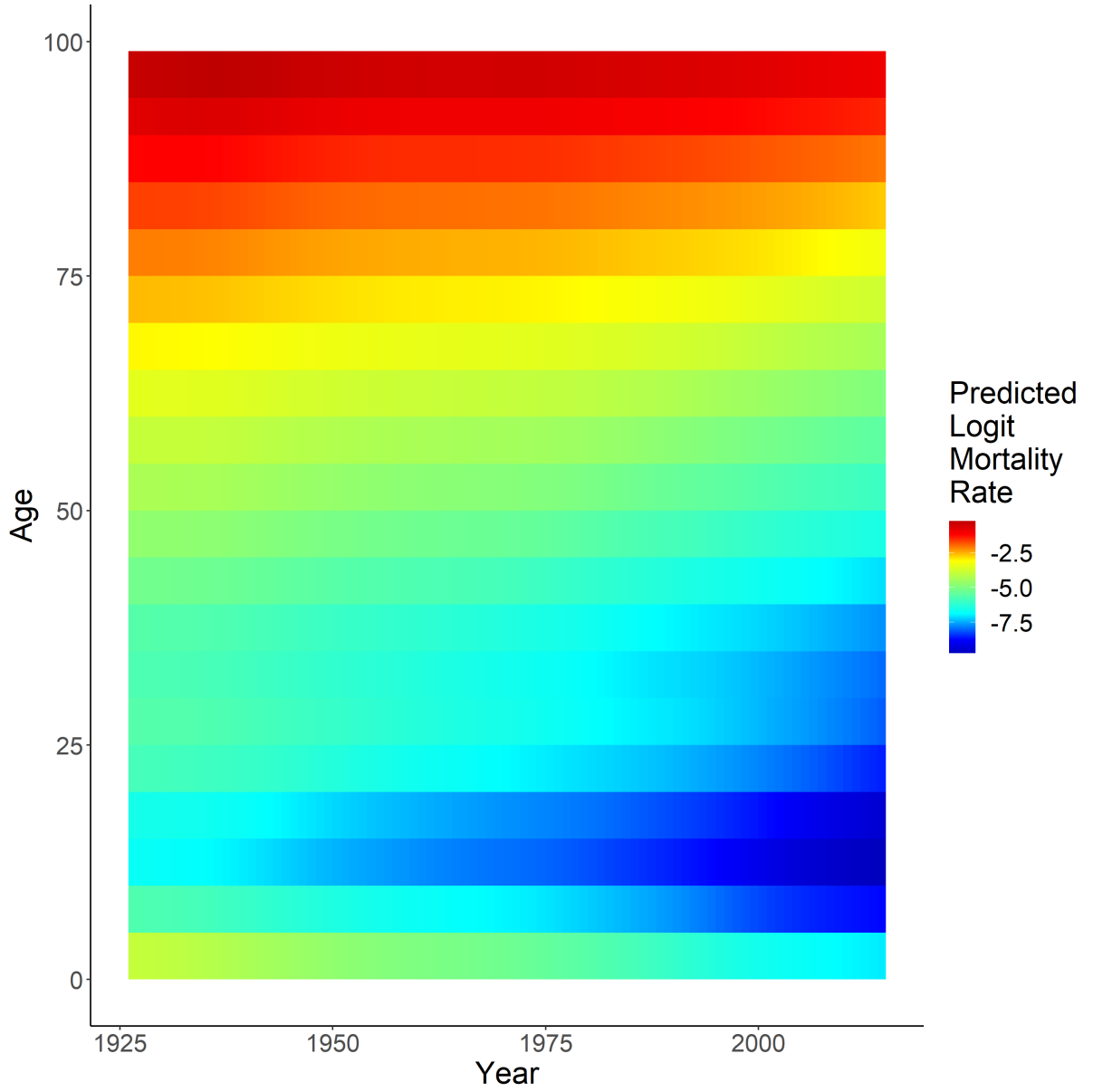


Figure A-15: Predicted all-cause mortality for years 1926-2015 and ages 0-99 in the United Kingdom for data aggregated in 5×1.

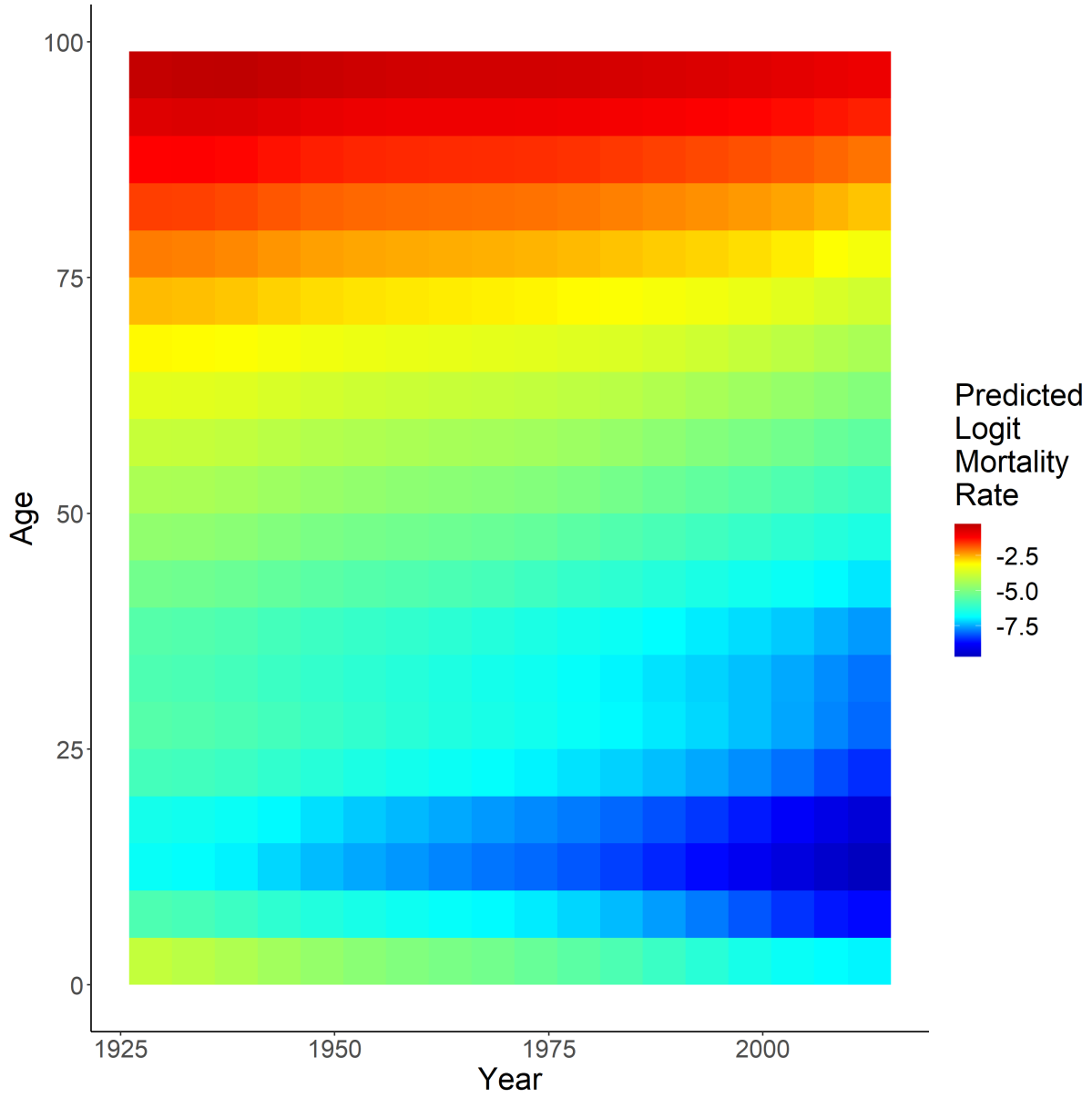


Figure A-16: Predicted all-cause mortality for years 1926-2015 and ages 0-99 in the United Kingdom for data aggregated in 5×5.

APPENDIX B

ADDITIONAL RESULTS FOR CHAPTER THREE

Here we present a number of additional results to supplement the results displayed in Chapter 3.

B.1 Additional material for the simulation study results

These are additional plots from the SU, RU and APC simulation study. In the SU and RU simulation study, only the age function was present in the data generation and in the APC simulation study, all three functions were present in the data generation. The estimates for each of the PSS and RW2 prior models are the average of each set of estimates over the full set of simulations. If the estimates fall on the $y = x$ axis, the estimates are equal.

B.1.1 Simple univariate model

Figure B-1 shows the comparison between the RW2 prior and PSS model estimates from the SU model simulation study. This plot shows how the RW2 prior and PSS model are equivalent to one another and can be used interchangeably for fitting a smooth function.

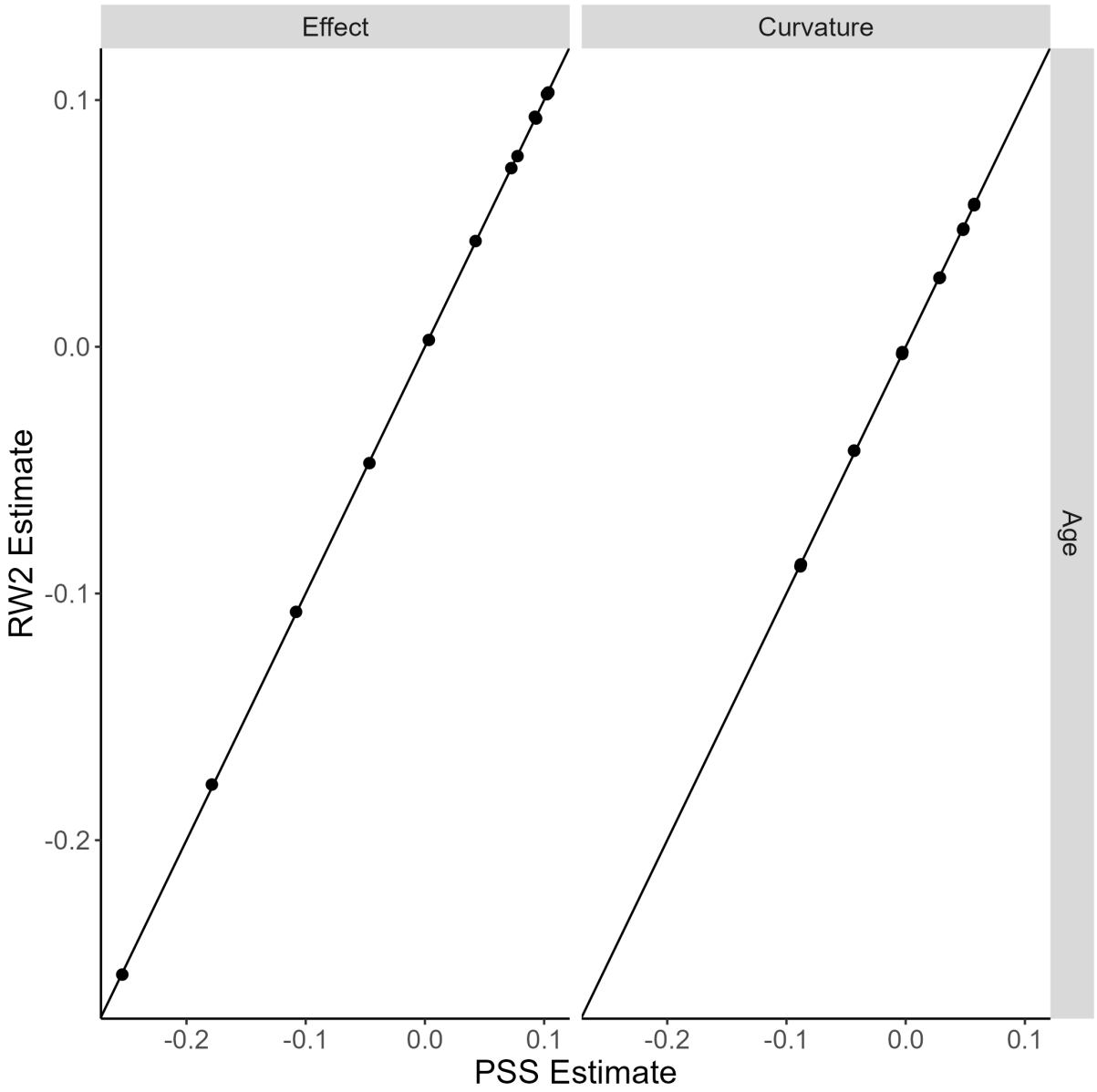


Figure B-1: Comparison between the RW2 prior and PSS estimates from the simple univariate simulation study.

B.1.2 Reparameterised univariate model

Figure B-2 shows the comparison between the RW2 prior and PSS model estimates from the RU model simulation study. This plot shows that the reparameterisation does not alter the correspondence and that it holds for fitting smooth functions of curvature.

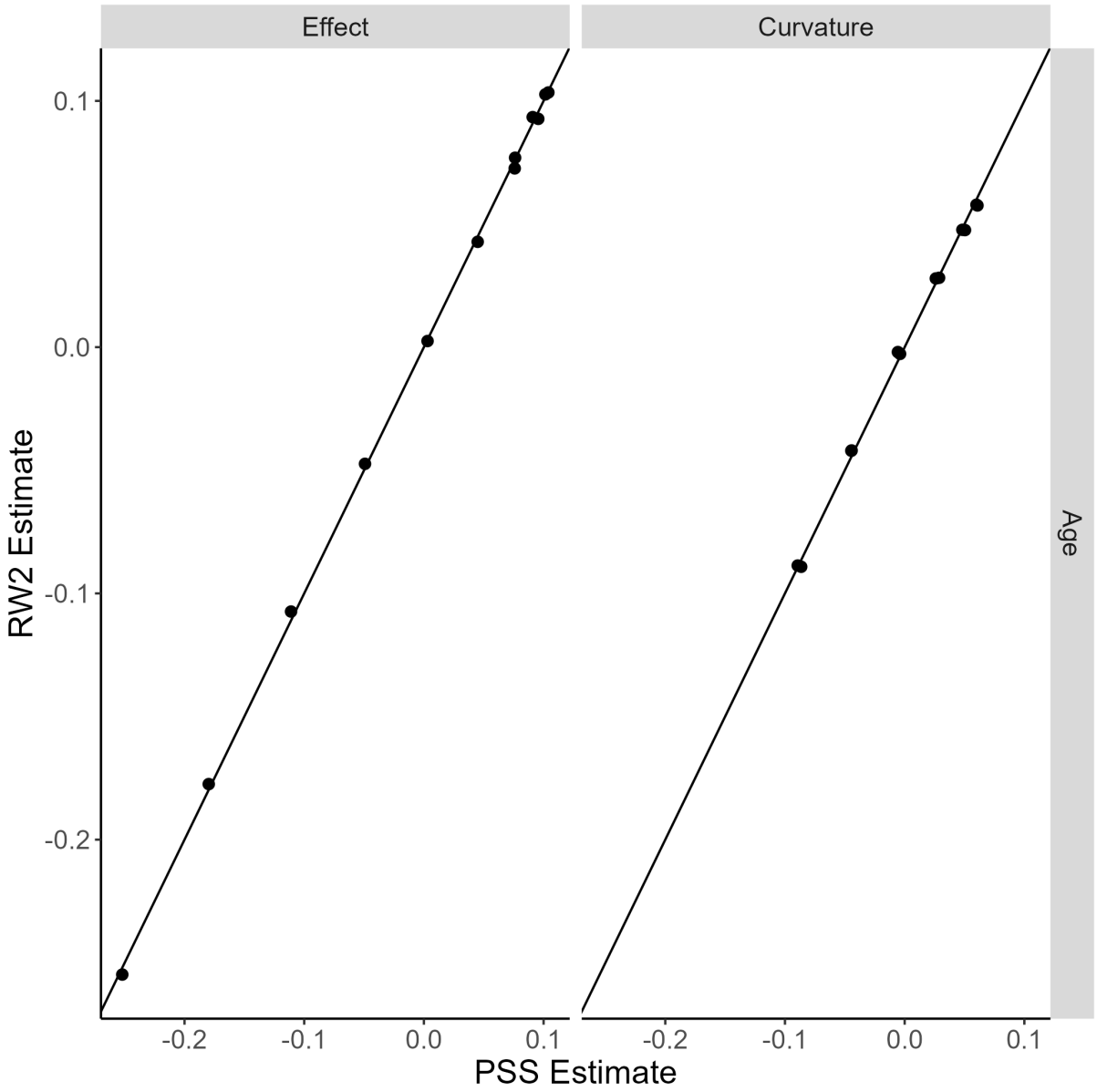


Figure B-2: Comparison between the RW2 prior and PSS estimates from the reparameterised univariate simulation study.

B.1.3 Age-period-cohort model

Figure B-3 shows the comparison between the RW2 prior and PSS model estimates from the APC model simulation study. The estimates of the full temporal effect not being on the $y = x$ axis is due to the structural link as these terms are not identifiable. The estimates of the curvature being on the $y = x$ axis show that the structural link and curvature identification problems do not alter the correspondence and that the RW2 prior model can be used interchangeably with the PSS model as an appropriate solution for fitting an APC model to data in unequal intervals.

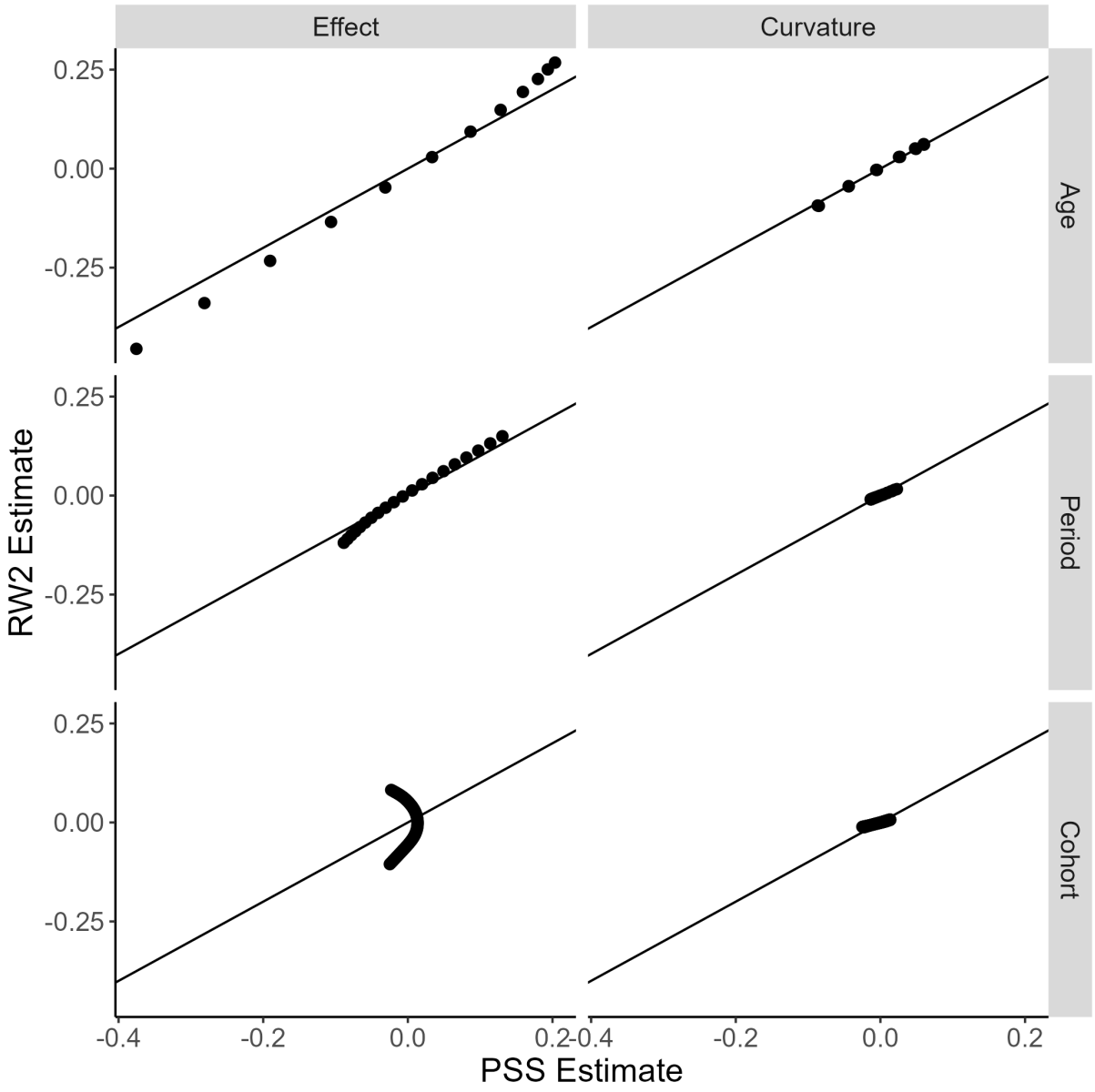


Figure B-3: Comparison between the RW2 prior and PSS estimates from the age-period-cohort simulation study.

APPENDIX C

ADDITIONAL RESULTS FOR CHAPTER FOUR

Here we include additional information and results to supplement the work in Chapter 4 including; details on model priors choices, posterior sampling, and pre-analysis conducted when defining the final model.

C.1 Proportions

Figure C-1 shows how the population proportion is split for the urban (top row) and rural (bottom row) strata in each of the region for the given years. Figure C-2 shows how the population is distributed across each region for the given years.

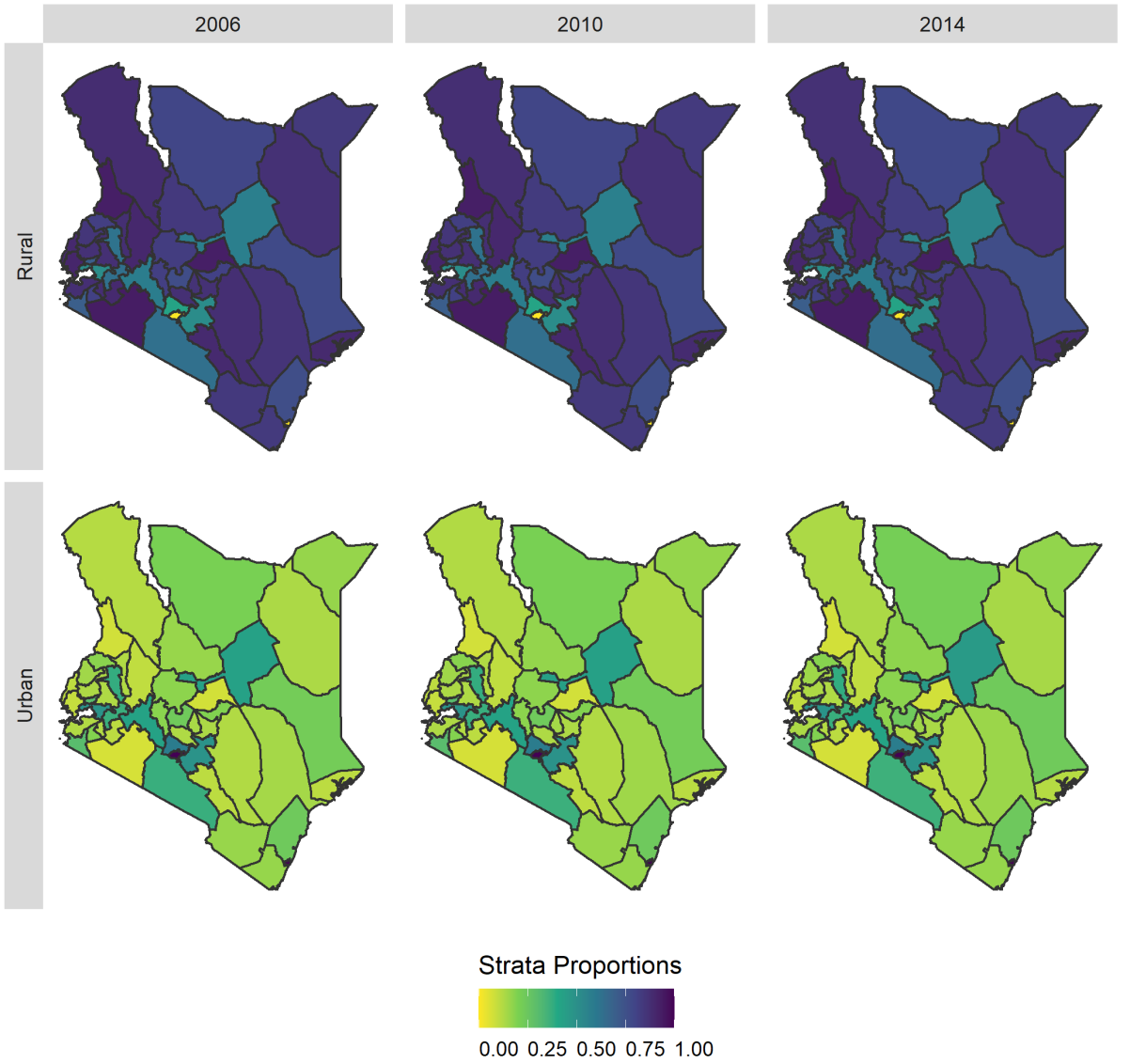


Figure C-1: Stratification proportions within each of the regions for 2006, 2010 and 2014. The top row are the rural proportions, and the bottom row is the urban proportions.

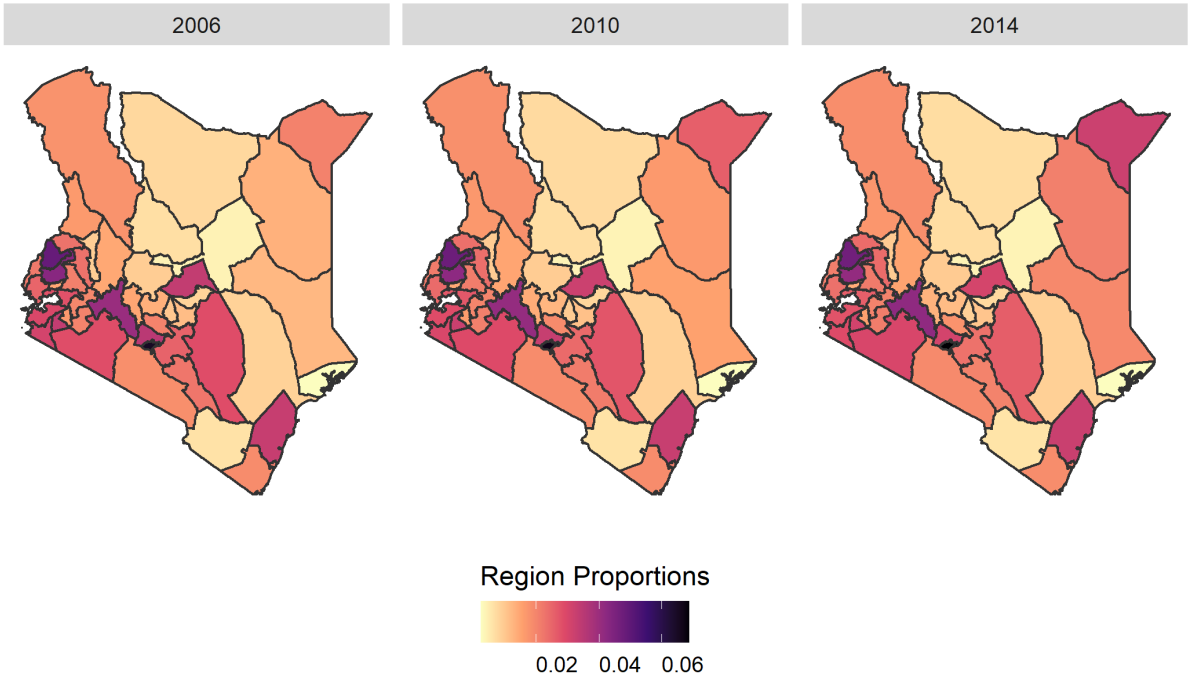


Figure C-2: Region proportions in Kenya for 2006, 2010 and 2014.

C.2 Priors

C.2.1 Period and cohort

We fit the model using a Bayesian hierarchical model in which prior distributions need to be assigned to all the random effect parameters in the model. To penalise deviations from the linear trend in a Bayesian sense, we use RW2 priors for each of the temporal curvature terms. For example, consider the period effects. The prior can be written,

$$\begin{aligned} f\left(\eta_p\right)|\tau_\eta &\propto \tau_\eta^{(P-2)/2}\exp\left(-\frac{\tau_\eta}{2}\sum_{p=1}^{P-2}\left(\eta_p - 2\eta_{p+1} + \eta_{p+2}\right)^2\right) \\ &= \tau_\eta^{(P-2)/2}\exp\left(-\frac{1}{2}\boldsymbol{\eta}_p^T\boldsymbol{Q}_\eta\boldsymbol{\eta}_p\right) \end{aligned}$$

with $\boldsymbol{Q}_\eta$, the precision matrix which depends on the unknown precision parameter τ_η, has rank $P-2$ and is of the form

$$\boldsymbol{Q} = \tau_\eta \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}. \tag{C.1}$$

The priors and precision matrices for the cohort curvature terms is analogous.

C.2.2 Non-constant intervals for age

Consider the age group midpoints being used in the KDHS, $a = 0.5, 6, 17.5, 29.5, 41.5, 53.5$. The difference between these is $5.5, 11.5, 12, 12, 12$ and 12, respectively. Since these are not all the same, the age model is not at regular time steps (intervals) and hence the random walk model used for period and cohort (that are at regular locations) does not apply to it; the precision matrix Eq.(C.1) is only defined at regular intervals.

When the data is at irregular locations, a computational efficient implementation of the RW2 model can be found by considering the RW2 model as the solution of a stochastic differential equation (SDE) which is found by using an Galerkin approximation (Lindgren and Rue, 2008). This method is an extension on the finite element method to approximate the solution of a

stochastic partial differential equations (SPDE) to fit continuous time models as implemented in `r-inla` (Lindgren et al., 2011). In addition, the Galerkin approximation to define the solution of the SDE is the default method of fitting a RW2 model (regardless of interval regularity) in `r-inla` as it is computationally efficient.

C.2.3 Spatial

Since we assume regions that border one another are more likely to be similar, we model the structured spatial component u_r with an intrinsic Gaussian Markov random field (Rue and Held, 2005) where the joint density is

$$\begin{aligned} f\left(u_r\right)|\tau_u &\propto \exp\left(-\frac{\tau_u}{2}\sum_{r\sim r'}\left(u_r-u_{r'}\right)^2\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{Q}_u\boldsymbol{u}\right). \end{aligned}$$

The notation $r \sim r'$ denotes that region r and r' are unordered neighbours and u_r has an intrinsic conditional autoregressive (ICAR) structure (Besag et al., 1991) where the precision matrix $\boldsymbol{Q}_u$ is of the form

$$\boldsymbol{Q}_u = \tau_u \begin{cases} n_r, & r = r' \\ -1, & r \sim r' \\ 0, & \text{else.} \end{cases}$$

Following Riebler et al. (2016), we use a BYM2 parameterization to model both the structured and unstructured spatial effects together,

$$S\left(\boldsymbol{s}_{r,k}\right) = \frac{1}{\tau_S}\left(\sqrt{\phi}u^*_{r[\boldsymbol{s}_k]} + \sqrt{1-\phi}v^*_{r[\boldsymbol{s}_k]}\right)$$

where τ_S controls the marginal variance controlling the weighted sum of $u^*_{r[\boldsymbol{s}_k]}$ and $v^*_{r[\boldsymbol{s}_k]}$, the structured and unstructured spatial effects. In addition, $\phi \in [0,1]$ is the mixing parameter that measures the proportion of the marginal variance explained by the structured spatial effect $u^*_{r[\boldsymbol{s}_k]}$.

C.2.4 Hyperpriors

For all the parameters we set the hyper-priors to be penalized complexity (PC) priors (Simpson et al., 2017). PC priors are designed to penalise deviations in the posterior distribution from a simpler, "base" model. PC priors were developed as a solution to the problem that arises when a prior forces overfitting (when it is impossible to distinguish between a model that is supported by the data or by a poor prior choice). For ζ, a flexibility parameter controlling how much the

model results depend on the data and the prior, $\mathcal{Q}(\zeta)$ is an interpretable transformation (i.e., standard deviation, precision) of ζ, U a "sensible" upper bound and p the probability of ζ being in the upper bound; a PC prior is specified

$$\mathrm{P}(\mathcal{Q}(\zeta) > U) = p.$$

The precisions for the temporal curvature RW2 terms have a PC prior where $U = 1$ and $p = 0.01$. For the BYM2 effect, the variance and scale parameters have $\mathrm{P}\left(\tau_S^{-1} > 1\right) = 0.01$ and $\mathrm{P}(\phi > 1/2) = 2/3$ PC priors, respectively. Finally, the space-time interaction term uses $U = 1/2$ and $p = 2/3$ for the joint space-time components.

C.3 Sampling from the posterior distribution of under-five mortality rates

The `r-inla` sampler will give draws from the posterior distribution (PD) of each of the terms (latent effects) in the linear predictor of Eq.(4.5) as well as giving the joint distributions of linear combinations of these latent effects. Therefore, we are able to define a PD of the monthly hazards,

$$\pi^{(w)}_{a[m],p,c,r} \sim p\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$$

where (w) is the w^{th} draw from the joint PD of the latent effects, $\boldsymbol{\theta}$ is the full set of parameters and $\boldsymbol{y}$ are the number of deaths.

Since the U5MR is a non-linear combination of samples, we are not able to sample from the U5MR PD from `r-inla` directly. However, we are able to define the U5MR PD by using each of the w draws from the monthly hazards PD in Eq.(4.6),

$$\text{U5MR}^{(w)}_{r,p} \sim p\left(\text{U5MR}_{r,p}\left(\boldsymbol{\theta}\right)|\boldsymbol{y}\right).$$

Explicitly, the U5MR depends on the parameters $\boldsymbol{\theta}$, but we can write the distribution without this dependence for simplicity, $\text{U5MR}^{(w)}_{r,p} \sim p\left(\text{U5MR}_{r,p}|\boldsymbol{y}\right)$.

C.4 Comparison of age-period model to SUMMER package version

Figures C-3 - C-4 show the comparison between two model that only have terms for age and period fit at the national level. The AT model is fit using the `SUMMER` package and the AP model is the sub-model of the APC model as described in Section 4.4.

These models are compared to see if reparameterisation is having any adverse effects. The similarity between the AT and AP models confirm this is not the case and the model is working as expected. There are slight differences between the lines in Figure C-3 but the large amount of overlapping between the 95% CI in Figure C-4 are an strong indication of the similarity between each of the methods.

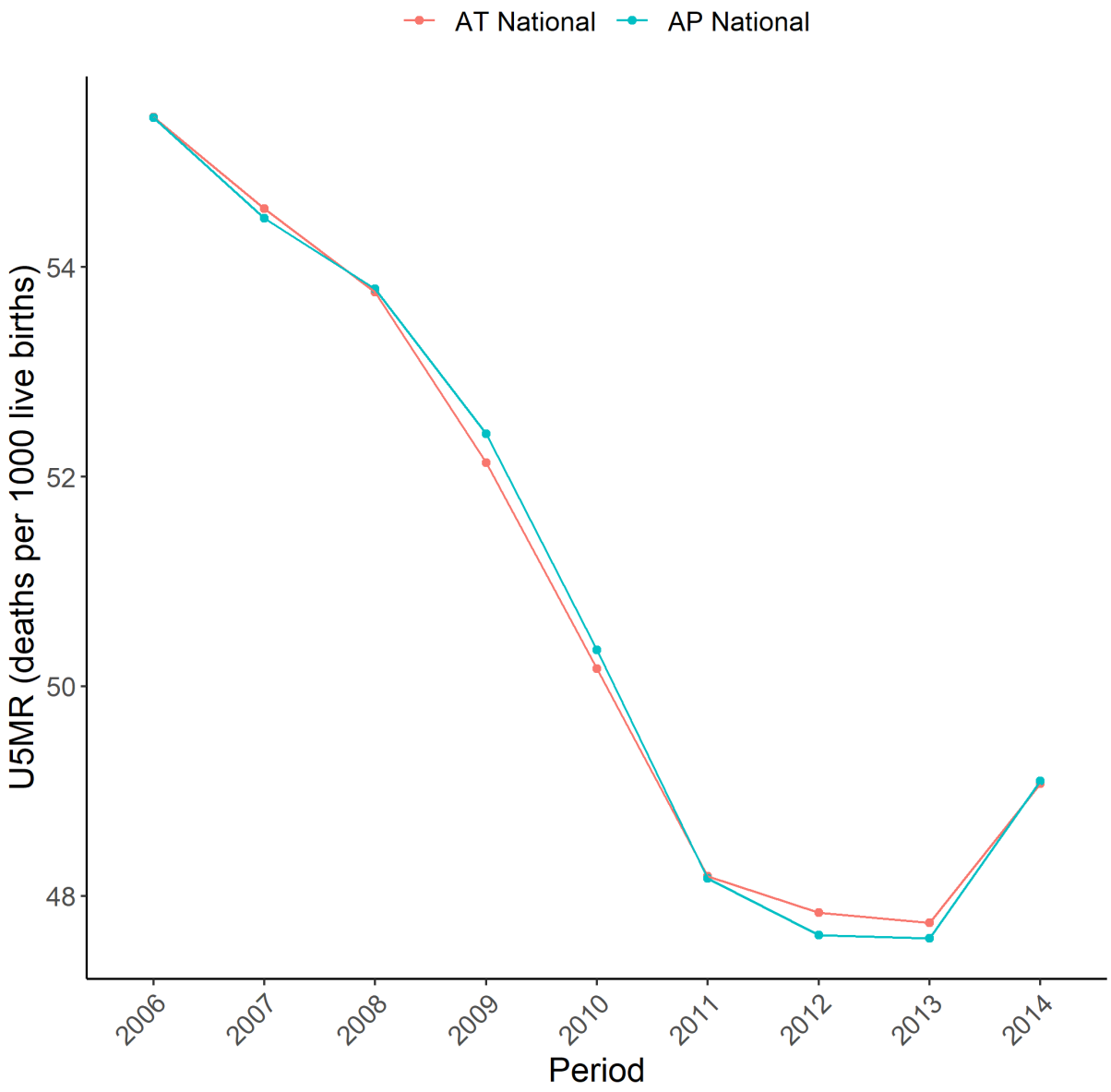


Figure C-3: AP and AT national estimates of U5MR for Kenya, 2006-2014.

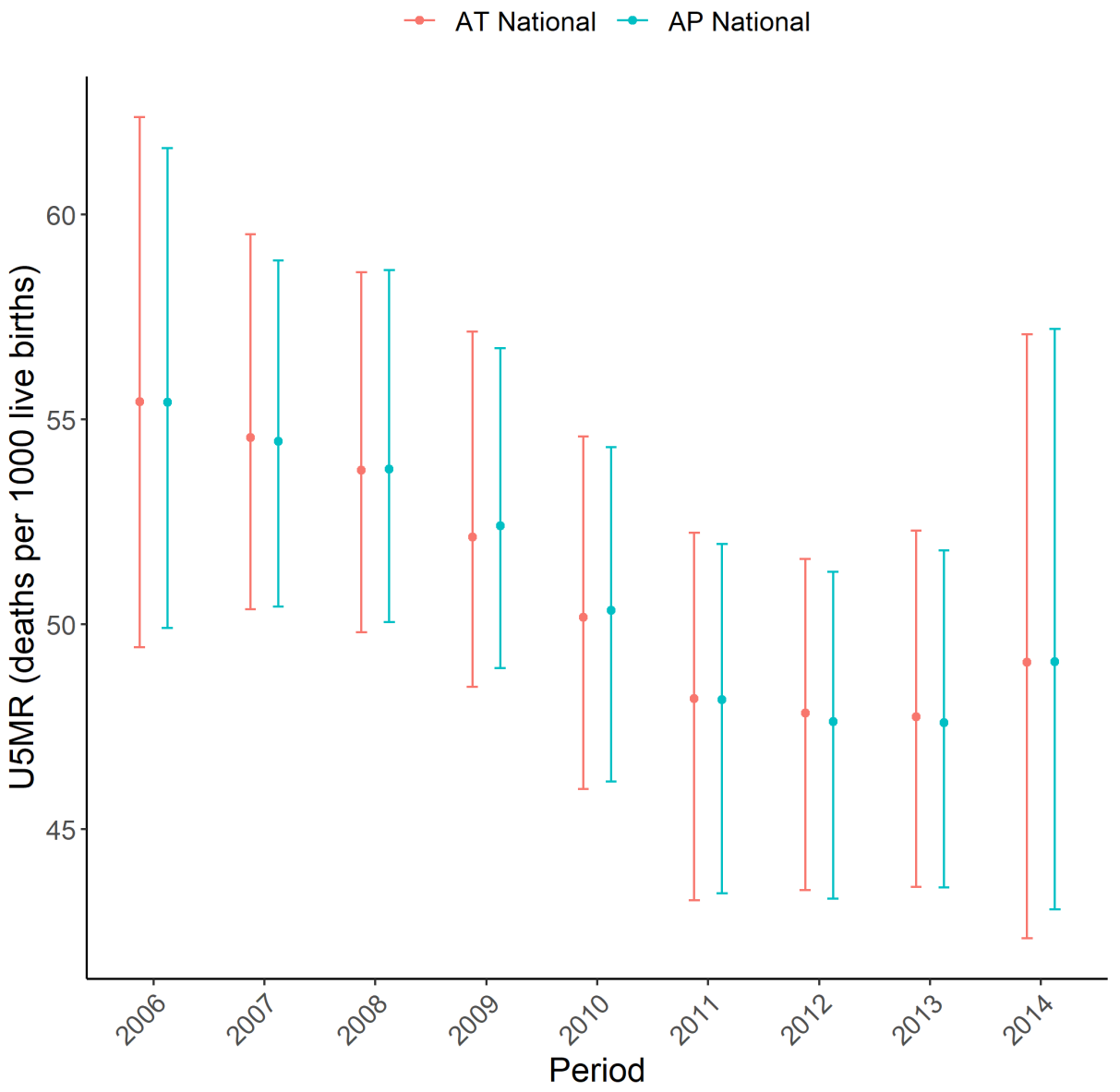


Figure C-4: AP and AT national estimates of U5MR for Kenya, 2006-2014 including 95% credible intervals.

C.5 Space-time interactions

C.5.1 Implementation

To define a space-time interaction (for all Type I-IV), first define a precision matrix for the spatial and temporal components $\boldsymbol{Q}_{\text{space}}$ and $\boldsymbol{Q}_{\text{time}}$, respectively; these can be either structured (such as a ICAR or RW2, respectively) or unstructured (iid). The precision matrix for the space-time interaction is the Kronecker product between these two precision matrices. The order of the Kronecker product when specifying the space-time precision matrix matters. In the DHS data, we believe the time order takes precedent over the space order, so the space-time precision is defined $\boldsymbol{Q}_{\text{space-time}} = \boldsymbol{Q}_{\text{time}} \otimes \boldsymbol{Q}_{\text{space}}$. To satisfy the linear constraints $\boldsymbol{Ax} = \boldsymbol{e}$ (Rue and Held, 2005), the additional constraints are defined via the eigenvectors that correspond to the zero eigenvalues of $\boldsymbol{Q}_{\text{space-time}}$. When implementing, all linear constraints must be ordered row-wise.

C.5.2 Model scores for different interaction types

Test	APC	AP	AC
DIC	**17267.83**	17338.32	*17344.31*
WAIC	**17267.87**	17339.19	*17344.90*

(a) Type I Interaction

Test	APC	AP	AC
DIC	**17267.43**	17338.12	*17343.93*
WAIC	**17267.61**	17338.26	*17344.01*

(b) Type II Interaction

Test	APC	AP	AC
DIC	**17267.66**	17338.34	*17344.48*
WAIC	**17267.87**	17338.49	*17344.41*

(c) Type III Interaction

Test	APC	AP	AC
DIC	**17266.89**	17338.53	*17342.01*
WAIC	**17267.38**	17338.56	*17342.34*

(d) Type IV Interaction

Table C.1: Model scores for each of an APC, AP, and AC model with the inclusion of a different space-time interaction. The best entries for each interaction type are in **bold**, whilst the worst entries are in *italics*. The best overall are highlighted in blue.

When comparing between which of the APC, AP, and AC model to include overall and what interaction type (Types I-IV) to include, we use the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and Watanabe-Akaike information criterion (WAIC) (Watanabe and Opper, 2010) which provide a measure of model fit and model complexity. Table C.1 shows the score for each model with each interaction. For each sub-table C.1a-C.1d, the values are bold is the lowest (best) and the values in italic are the highest (worst). In addition, the values highlighted in blue are the lowest overall.

For each interaction, the APC model is the best fitting model of the three with the AC model being the worst. The Type IV interaction model is the best for APC and AC, but the Type II model is the best for the AP model. When choosing between what temporal terms to include

and what type of interaction term, the choice of the temporal term is far more influential on the model fit. Whilst the DIC and WAIC score do not have units, the relative comparisons highlight the difference. Within each interaction model, the difference between the temporal models is far more than the choice between what interaction term to include within each model. Given the results in Table C.1, the difference between interaction types is negligible so we will consider a Type IV interaction within the main paper. We do however include the results for each of the temporal models (APC, AP, and AC) as the difference between these is far more noticeable.

C.5.3 Establishing space-time interactions

Figures C-5 - C-8 shows the line plots of U5MR per 1000 live births for each of the 47 regions in each space-time interaction type. If there was no space-time interaction, each of the lines in the plots would be parallel to one another. Since there are multiple overlapping lines in all four plots, there is a clear space-time interaction effect occurring for all the interaction types being considered.

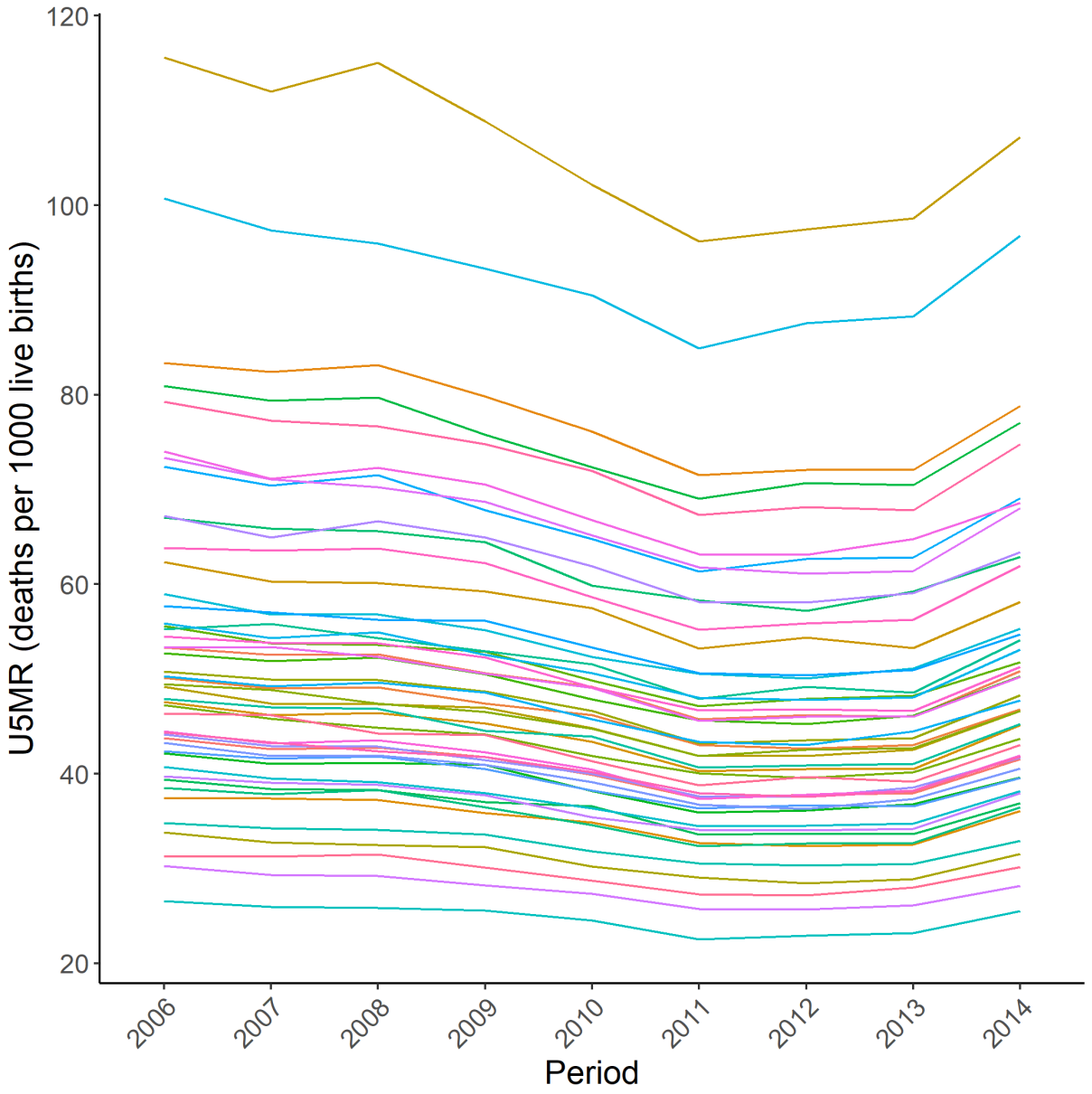


Figure C-5: Line plot of U5MR per 1000 live births for individual regions in Kenya, 2006-2014, for a Type I Interaction.

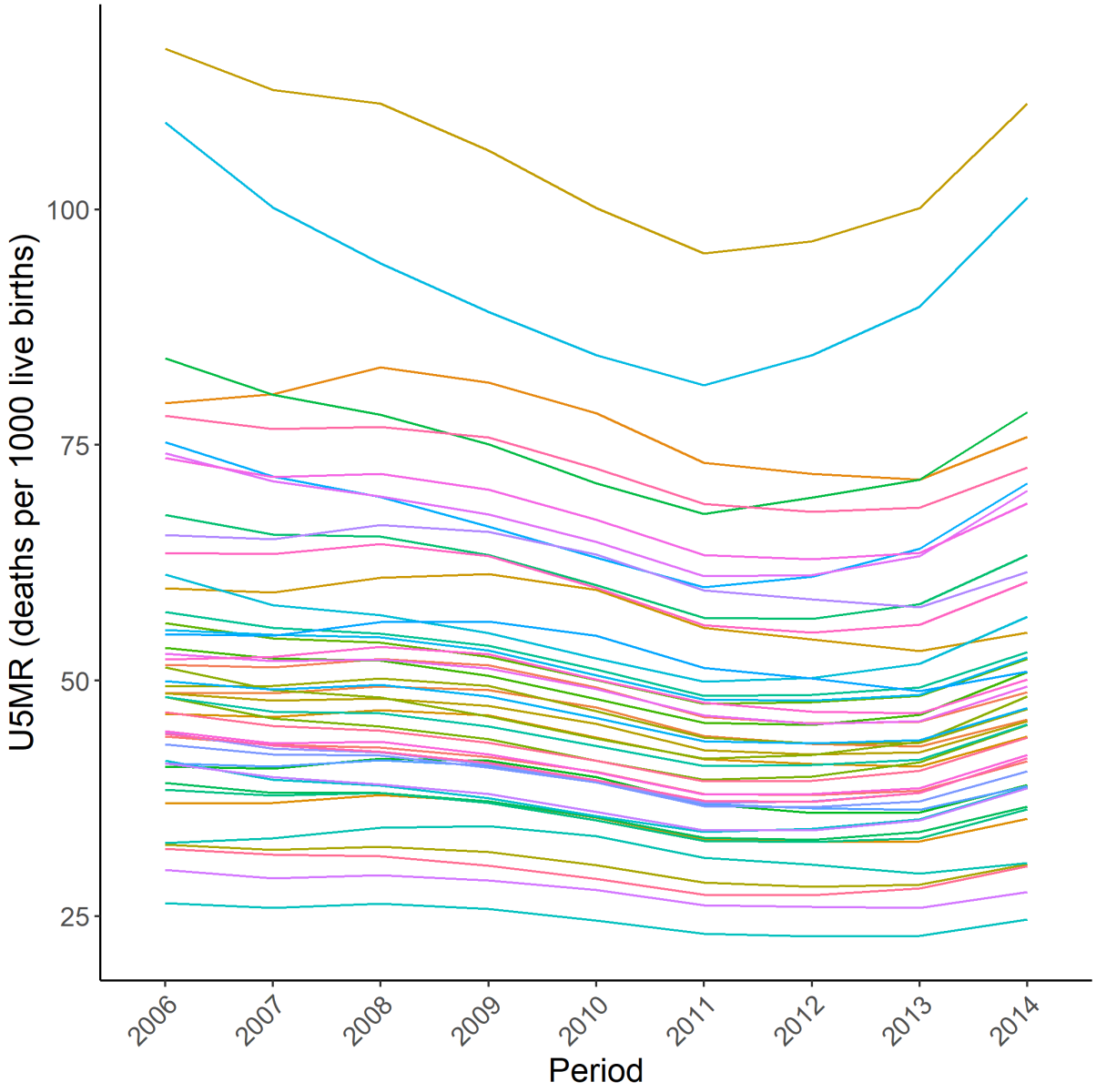


Figure C-6: Line plot of U5MR per 1000 live births for individual regions in Kenya, 2006-2014, for a Type II Interaction.

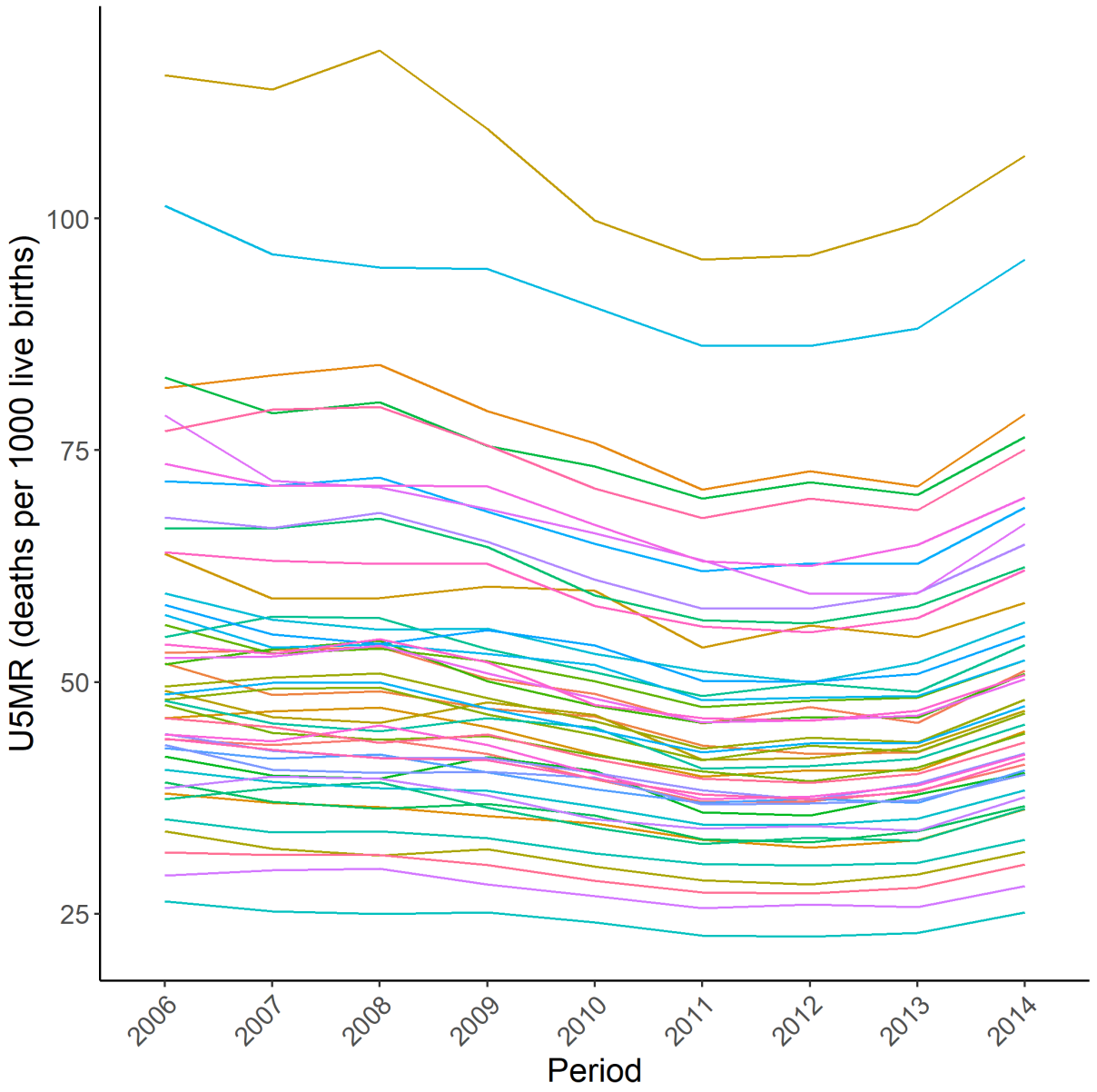


Figure C-7: Line plot of U5MR per 1000 live births for individual regions in Kenya, 2006-2014, for a Type III Interaction.

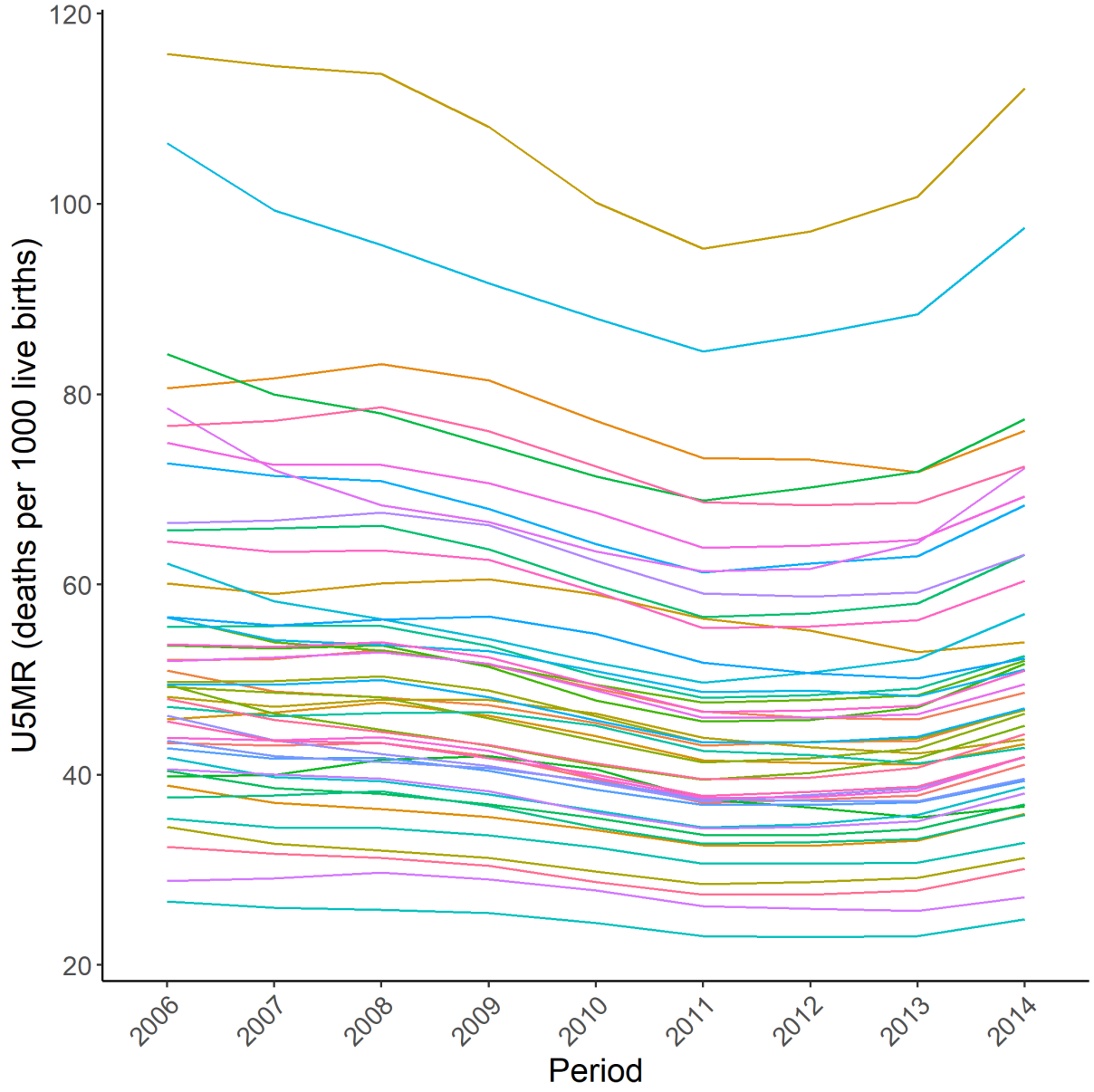
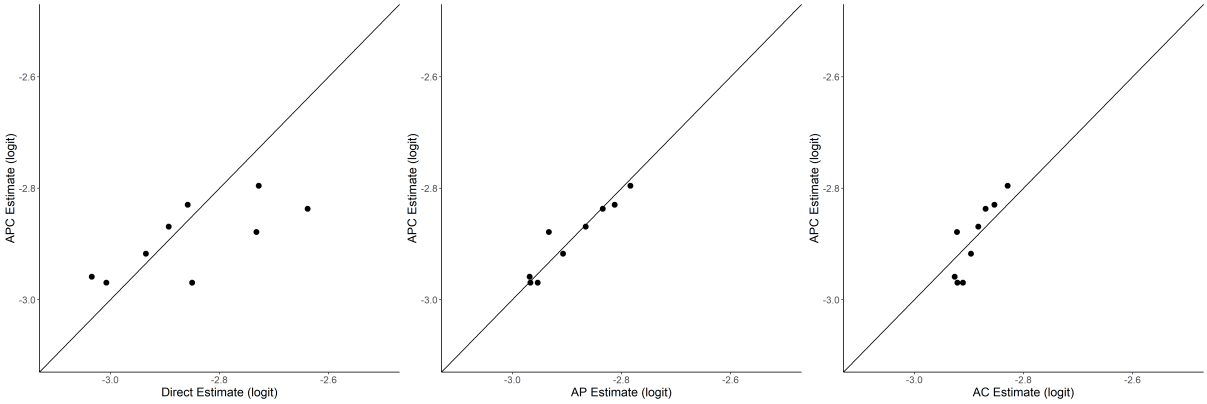


Figure C-8: Line plot of U5MR per 1000 live births for individual regions in Kenya, 2006-2014, for a Type IV Interaction.

C.6 Yearly, national age-period-cohort estimates of under-five mortality rates verses other estimates

Figure C-9 show the yearly, national APC model estimates compared against those for the direct, AP and AC (left-to-right) estimates on the logit scale. The attenuation of the APC, AP and AC models is expected. In addition, the results from Figures 4-2 and 4-3 mean we expect the closeness between the APC estimates and both the AP and AC estimates. Whilst the APC estimates are not as close to the direct estimates as they are to the AP and AC estimates, they are still relatively similar. Therefore, the APC model is performing well when compared to the gold-standard direct estimates.



(a) Versus APC estimates. (b) Versus AP estimates. (c) Versus AC estimates.

Figure C-9: Yearly, national-level direct, AP and AC estimates versus APC estimates on the logit scale. From left-to-right are the direct, AP and AC models, respectively.

C.7 Validation

C.7.1 Sampling Distribution for under-five mortality rates

In the leave-one-out (LOO) cross validation (CV), we systematically leave out observations (deaths) and use the remaining observations to predict over those missing. We will systematically leave out all observations in the period 2014 one region at a time and use the remaining observations to predict the missing values (hazards). Let $\boldsymbol{y}_{-p=2014,r}$ represent the all the observations where those from period 2014 and region r missing; implicitly, we are leaving out all ages $a\,[m]$ and all cohorts c. For brevity, we further simplify this to $\boldsymbol{y}_{-r}$, since $p = 2014$ is fixed.

Using the sampler from `r-inla` and the U5MR formula, Eq.(4.6), we predict the missing observations from the remaining data leaving the following PD for the U5MR,

$$\text{U5MR}_r^{(w)} \sim p\left(\text{U5MR}_r|\boldsymbol{y}\right)$$

where for simplicity, have dropped p from the subscripts since $p = 2014$ is fixed. In addition, (w) denotes the w^{th} draw from the PD and we take W total draws from the PD to incorporate the sampler variability fully. In order to incorporate the variability from the complex survey design in the predictions, we use the sampling distribution (SD) from Mercer et al. (2014),

$$\widetilde{Y}_r^{(w),(m)} \sim \text{N}\left(Y_r^{(w)}, \hat{V}_{\text{r}}^{\text{Des}}\right)$$

to generate M samples for each of the W samples from the PD. In doing this, we are fully considering the design variability for each of the W samples. As in Chapter 4, $Y_{r,p} = \text{logit}\left(\text{U5MR}_{r,p}\right)$ and we can drop the p subscript since $p = 2014$ is fixed. The sampler variability is incorporated through $Y_{r,p}$ and the complex design variability is incorporated through $\hat{V}_{\text{r}}^{\text{Des}}$, the variance of the direct estimate for region r in 2014.

We have R regions which we generate W samples for from the PD. For each of the W samples, we generate M samples from the SD using $Y_r^{(w)}$ as the mean. Therefore, we have an $R \times W \times M$ array for the full SD which incorporates both the sampling and complex design variance in estimates for U5MR for the validation scores. Figure C-10 shows how the sampling array looks for general R, W and M. In order to have a final estimate, we take an global average over both the W and M samples from each of the PD and SDs. Therefore, the explicit final estimate for the U5MR that is used in the score is,

$$\overline{\widetilde{Y}_r} = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{W}\sum_{w=1}^{W}\widetilde{Y}_r^{(w),(m)}\right]$$

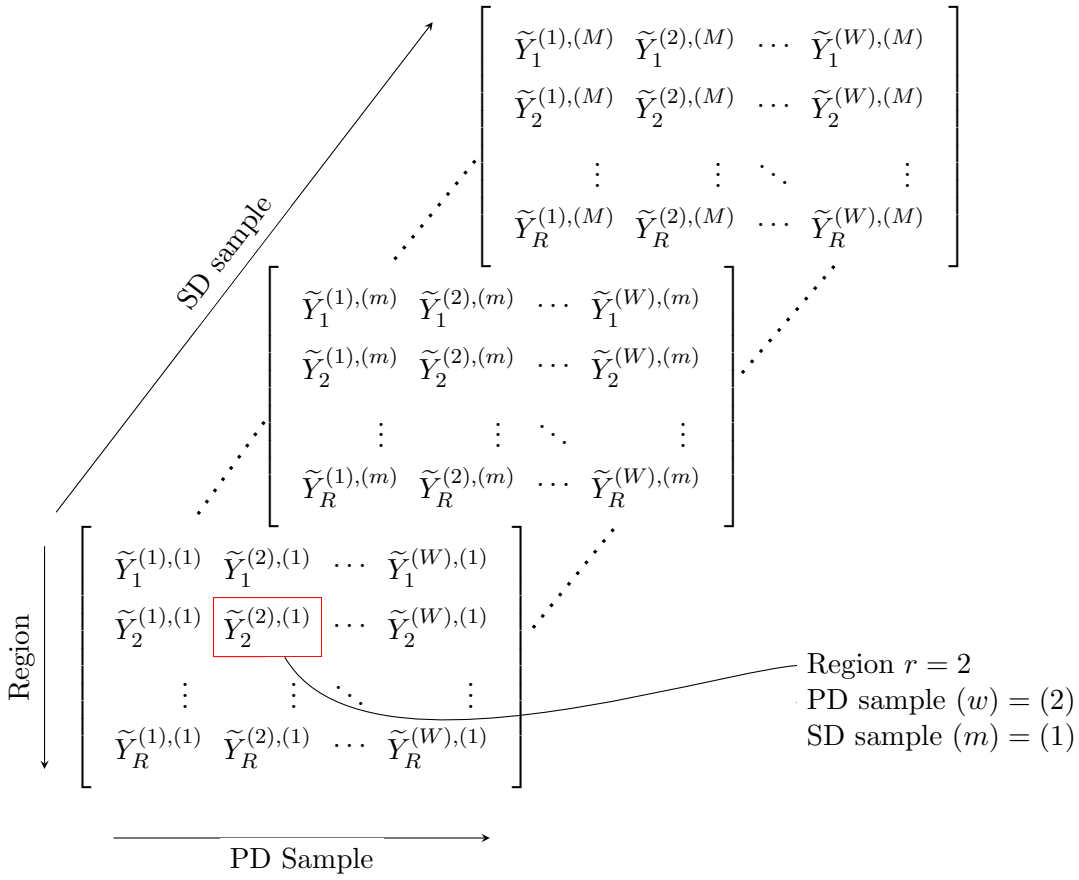


Figure C-10: Array of the samples from both the posterior and sampling distributions of U5MR to incorporate both the sampler and design variability in the under-five mortality rate on the logit scale.

C.7.2 Predicted subnational maps

Figure C-11 is the predicted U5MR per 1000 live births and Figure C-12 is the width of the 95% CI. Both plots here are like the equivalent maps for the 2014 estimates in Figures 4-4 and 4-5 which indicates how well the APC model is at predicting. The preserved spatial structure is shown by areas near to one another having similar U5MR and the wider 95% CI correspond to the regions with higher U5MR. For example, the regions in the west of Kenya have both a higher U5MR per 1000 live births and a higher 95% CI width.

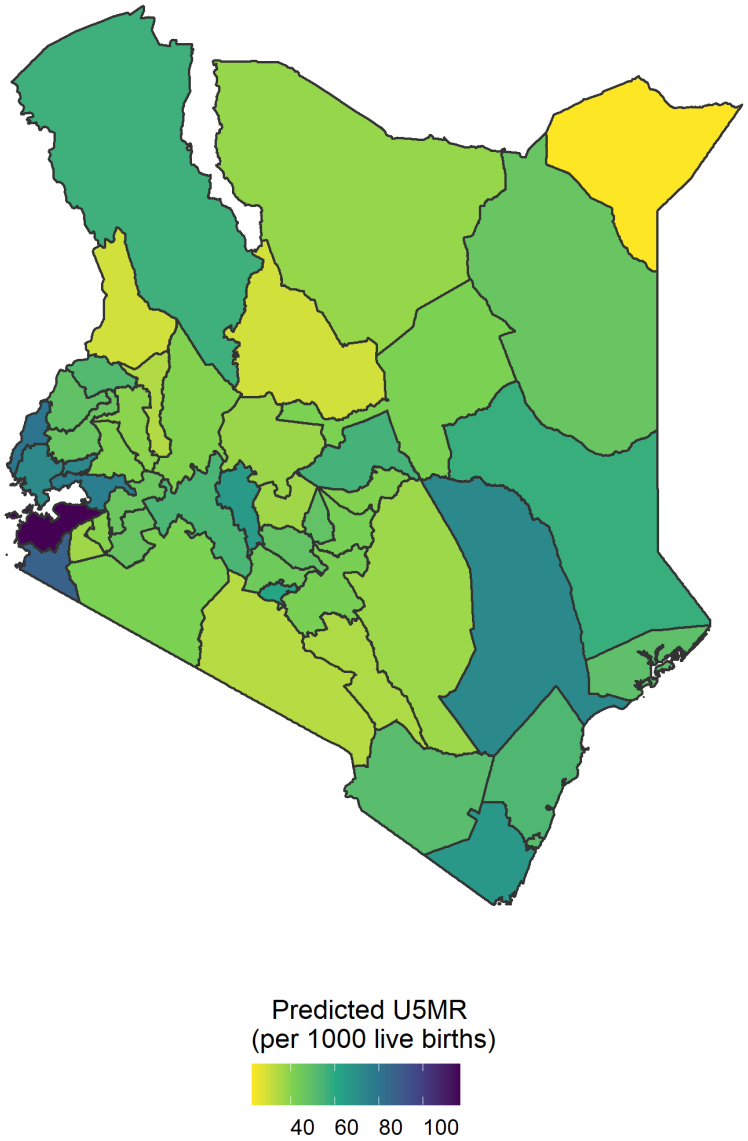


Figure C-11: Predicted under five mortality rates for 2014 from the LOO CV.

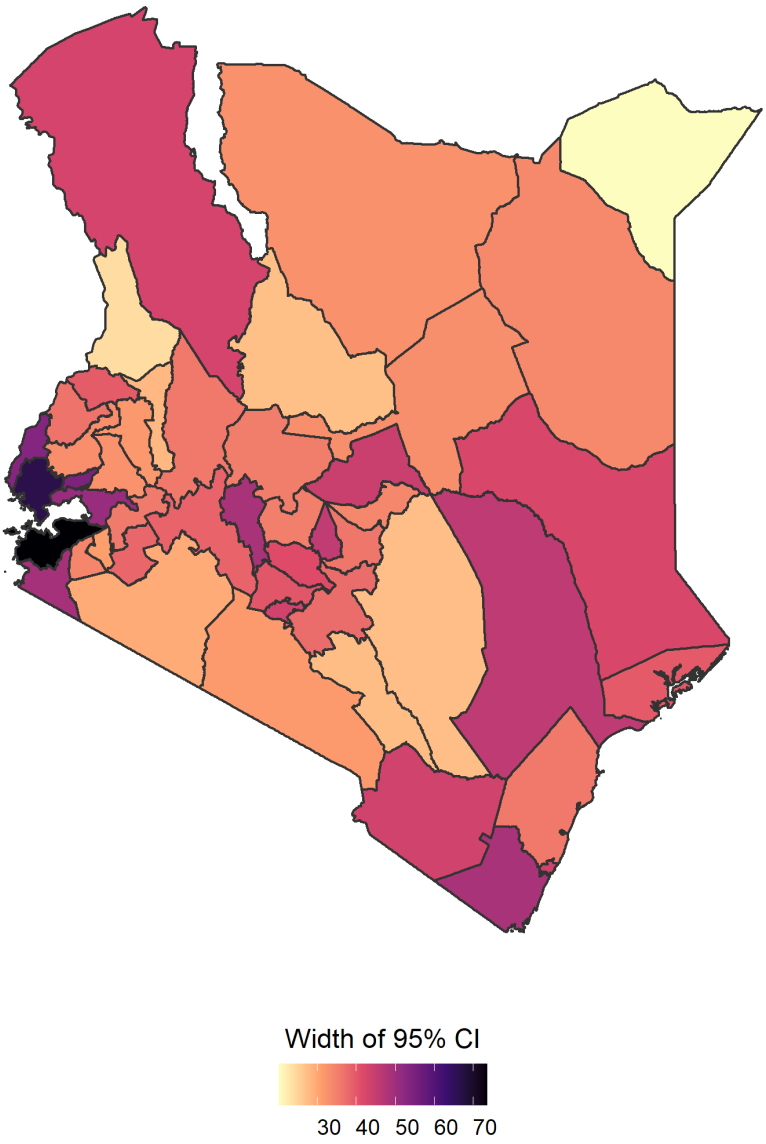


Figure C-12: Width of 95% credible intervals for 2014 from the LOO CV.