Comparing Temporal Smoothers for use in Demographic Estimation and Projection*

Monica Alexander†
University of California, Berkeley

October 31, 2017

Abstract

The development of methods to estimate and project demographic and health indicators is important to help monitor trends over time. In practice, estimation often occurs in situations where data are sparse or variability is high. Trends and projections may be unclear because of missing observations over time, or if the observed data do not follow a smooth trajectory. Determining how data observations should be modeled and smoothed over time is not always a straightforward process. The aim of this paper is to compare the characteristics and performance of different temporal smoothing techniques to gain a deeper understanding into which methods work well in different data availability situations and how sensitive the resulting estimates are to modeling decisions. A review of the three modeling families (ARMA models, Gaussian process regression, and penalized splines regression) is presented, highlighting the main similarities and differences across the methods. Model performance is evaluated on both simulated and real data, focusing on two common data scenarios: small populations; and data-sparse situations.

1 Introduction

Accurate measurement of demographic indicators over time is important for monitoring progress at the regional, national and subnational level. Examples of such indicators include all-cause child or maternal mortality, cause-specific mortality, fertility rates and contraceptive prevalence, or unmet need in contraceptive use. To effectively track trends and progress in such indicators, statistical models are often employed to obtain estimates that are as accurate and reliable as possible, to project trends into the future, and to get a sense of the uncertainty around these estimates and projections.

*I would like to thank Leontine Alkema for helpful suggestions and comments on this paper. This work was partially funded by the Department of Health Statistics and Information Systems at the World Health Organization.

†monicaalexander@berkeley.edu

In practice, estimation often occurs in situations where data are sparse or variability is high. Trends and projections may be unclear because of missing observations over time, or if the observed data do not follow a smooth trajectory. Determining how data observations should be modeled and smoothed over time is not always a straightforward process.

Model frameworks to estimate time series of demographic indicators commonly consist of two main parts. The first part is a regression model that expresses the expected level and trend of the outcome based on some related covariates. The second part is a temporal smoothing process that allows for non-linearities in the data to be captured over time. In addition, the temporal model explicitly allows for the outcome to be forecast and uncertainty intervals to be produced.

A survey of the literature suggests the temporal smoothing component of demographic estimation models is usually one of three families: ARMA (time series) models, Gaussian process regression, and penalized splines regression. For example, first- and second-order autoregressive (AR) processes are used in models to estimate contraceptive prevalence (Alkema et al., 2013) and blood pressure (Finucane et al., 2014) in countries worldwide. An autoregressive-moving-average model is used in the estimation of maternal mortality in all UN-member countries. Penalized splines regression has been used to estimate and project child mortality (Alkema and New (2014); Alexander and Alkema (2016)) and adult mortality (Currie et al., 2004). Gaussian process regression has also been used in many contexts, including child mortality and cause-specific mortality (Foreman et al., 2012). While the technique chosen in each case appears to perform well, it is not always clear why one temporal smoothing technique was chosen over another, and how sensitive the model results would be to different decisions.

The aim of this paper is to compare the characteristics and performance of these different temporal smoothing techniques to gain a deeper understanding into which methods work well in different data availability situations and how sensitive the resulting estimates are to modeling decisions. A review of the three modeling families is presented, highlighting the main similarities and differences across the methods. Model performance is evaluated on both simulated and real data, focusing on two common data scenarios: small populations; and data-sparse situations. The paper concludes with a discussion about implications for thinking about uncertainty and model choice.

2 Methods

2.1 Formulation of general modeling framework

Consider the situation of estimating and projecting an outcome over time. This quantity could be an indicator such as the infant mortality rate, the lung cancer mortality rate, or the proportion of women using some form of contraception. It is often the case that models for these outcomes include one or more covariates that are known to be related in a systematic way. For example, a model used by the World Health Organization for estimating maternal mortality rates for all UN-member countries assumes maternal mortality is a function of GDP, the fertility rate and percent of skilled attendants at birth (Alkema et al., 2016). However, often models that only include covariates cannot adequately capture temporal trends observed

in the data. As such, non-linear temporal smoothing methods are added to the underlying covariate model.

Continuing with the maternal mortality example, data-driven trends are modeled through the inclusion of a time series model that captures accelerations and decelerations in the rate of change in the maternal mortality. This general modeling approach, where an outcome of interest is modeled as a combination of an expected level given covariates and distortions around this expected trend, has been used in many different scenarios.

Formally, define θ_t to be the quantity of interest at time t in a particular area. Define an additive model for θ_t of the form:

$$\theta_t = \psi_t + X_t + \varepsilon_t, \tag{1}$$

where ψ_t is the expected level of θ_t given covariates, X_t are distortions away from this expected level at time t and ε_t is an error term.

The focus of this paper is considering different ways to model of the distortions, X_t. Of course, the choice of how to model the expected level, ψ_t, is also important and can affect the resulting estimates substantially. However, in general there has been less discussion and illustrations in the literature of sensitivities to the choice of temporal smoothing method for X_t.

2.2 Summary of three main modeling families

Three main method families are considered to model X_t: time series (ARMA) models; Gaussian process regression; and penalized splines regression. Their main characteristics are explained below, and then similarities and differences between the methods are discussed in the next section.

2.3 Time series (ARMA) models

Autoregressive moving average (ARMA) models are fitted to time series data, allowing for autocorrelation (correlation through time) to be taken into account (Box et al., 2015). The autoregressive (AR) part assumes that the variable of interest is dependent on its past values. The moving average (MA) part assumes the error in the regression can be expressed as a linear combination of past errors.

ARMA models are described as ARMA(p,q) where parameters p, and q refer to the order (number of time lags) of the AR model and the order of the MA model, respectively. In demographic applications, ARMA models usually have relatively low orders (two or less). This paper focuses on AR(1) and ARMA(1,1) models.

A first-order Autoregressive process, or AR(1), can be written as

$$\begin{aligned} X_t &= \rho X_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim N(0, \sigma^2). \end{aligned}$$

This implies that an observation at time t is dependent on the previous observation, plus some error. The larger the autoregressive coefficient, ρ, the greater the covariance through time.

First-order Autoregressive Moving Average models, i.e. ARMA(1,1) are like AR(1) but include an additional term that allows errors at a particular time t to

be dependent on the previous error:

$$\begin{aligned} X_t &= \rho X_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim N(0, \sigma^2). \end{aligned}$$

These processes are stationary, that is, never diverge from fluctuating around zero, if $|\rho| < 1$. For a stationary series, the covariance structure as described by the covariance function, $k(t, t+s)$, is independent of t, i.e. $k(t, t+s)$ depends only on the distance s. The covariance between points at time t and $t+s$ can be expressed as a function of the model parameters, ρ, σ (and θ for ARMA(1,1)). In particular, the covariance between X_t and X_{t+s} in an AR(1) process is

$$k(t, t+s) = \frac{\sigma^2}{1-\rho^2} \cdot \rho^{|s|}. \tag{2}$$

For ARMA(1,1), the stationary covariance between X_t and X_{t+s} is

$$k(t, t+s) = \frac{\sigma^2(\rho+\theta)(1+\rho\cdot\theta)}{1-\rho^2} \cdot \rho^{|s|}. \tag{3}$$

Figure 1 illustrates example AR(1) and ARMA(1,1) processes. The larger the value for ρ, the higher the autocorrelation and the more regular the pattern. For the ARMA(1,1) process, the pattern is also affected by the value for the MA term, θ. As θ approaches zero, ARMA(1,1) approaches an AR(1) process.

An extension of ARMA models is ARIMA models, where the I stands for 'integrated' and refers to whether the analysis is performed on a differenced series. This is undertaken if the series on the original scale is not stationary.

2.4 Gaussian process regression

Gaussian processes extend multivariate Gaussian (Normal) distributions to infinite dimensionality. They are a collection of data points such that any finite subset of the range follows a multivariate Gaussian distribution. Gaussian processes are defined in terms of a mean and variance-covariance function, and can form the basis of a regression to estimate and predict new data points.

Formally, for any sequence of times, $\boldsymbol{t} = t_1, t_2, \ldots, t_n$ a Gaussian process (GP) can be written as

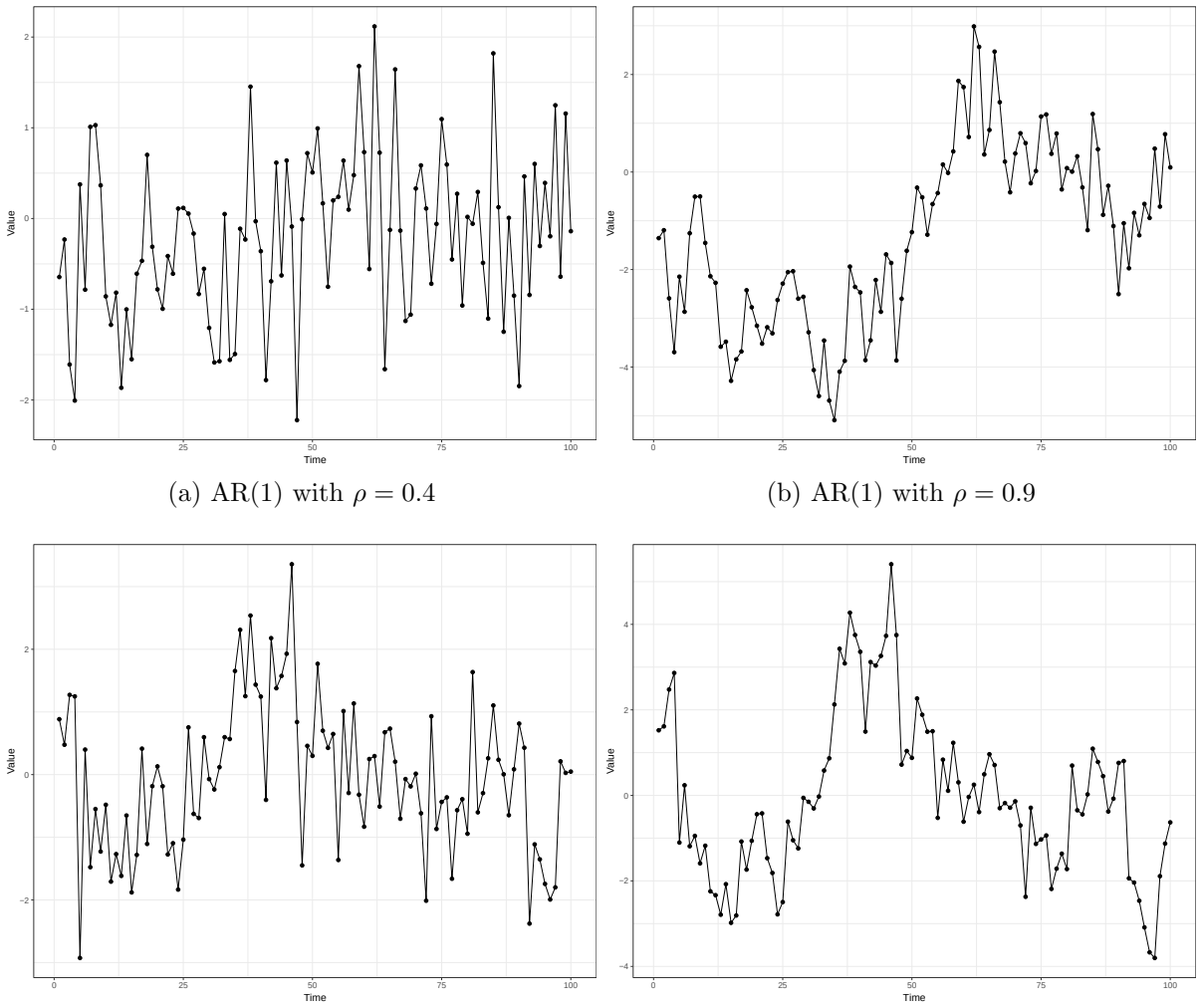
$$X_{\boldsymbol{t}} \sim GP\left(m(\boldsymbol{t}), k(\boldsymbol{t}, \boldsymbol{t}')\right).$$

with mean function $m(\boldsymbol{t})$ and covariance function $k(\boldsymbol{t}, \boldsymbol{t})$. Usually the mean function is set to be zero, as any systematic mean trends are modeled through the expected level (through the ψ_t term in Equation 1).

The covariance function can be chosen from a wide range of appropriate functions. This paper focuses on two choices, the squared exponential and Matern functions, which have been used in demographic modeling. This squared exponential covariance function is expressed as

$$k(t, t+s) = \sigma^2 \exp\left(-\frac{|s|^2}{2\lambda^2}\right). \tag{4}$$

The λ parameter is a length parameter which controls the smoothness of the fit of the Gaussian process. The smaller the value λ, the larger the covariance and



(a) AR(1) with $\rho = 0.4$

(b) AR(1) with $\rho = 0.9$

(c) ARMA(1,1) with $\rho = 0.9$ and $\theta = 0.6$

(d) ARMA(1,1) with $\rho = 0.9$ and $\theta = 0.1$

Figure 1: Example AR(1) and ARMA(1,1) with different parameter values

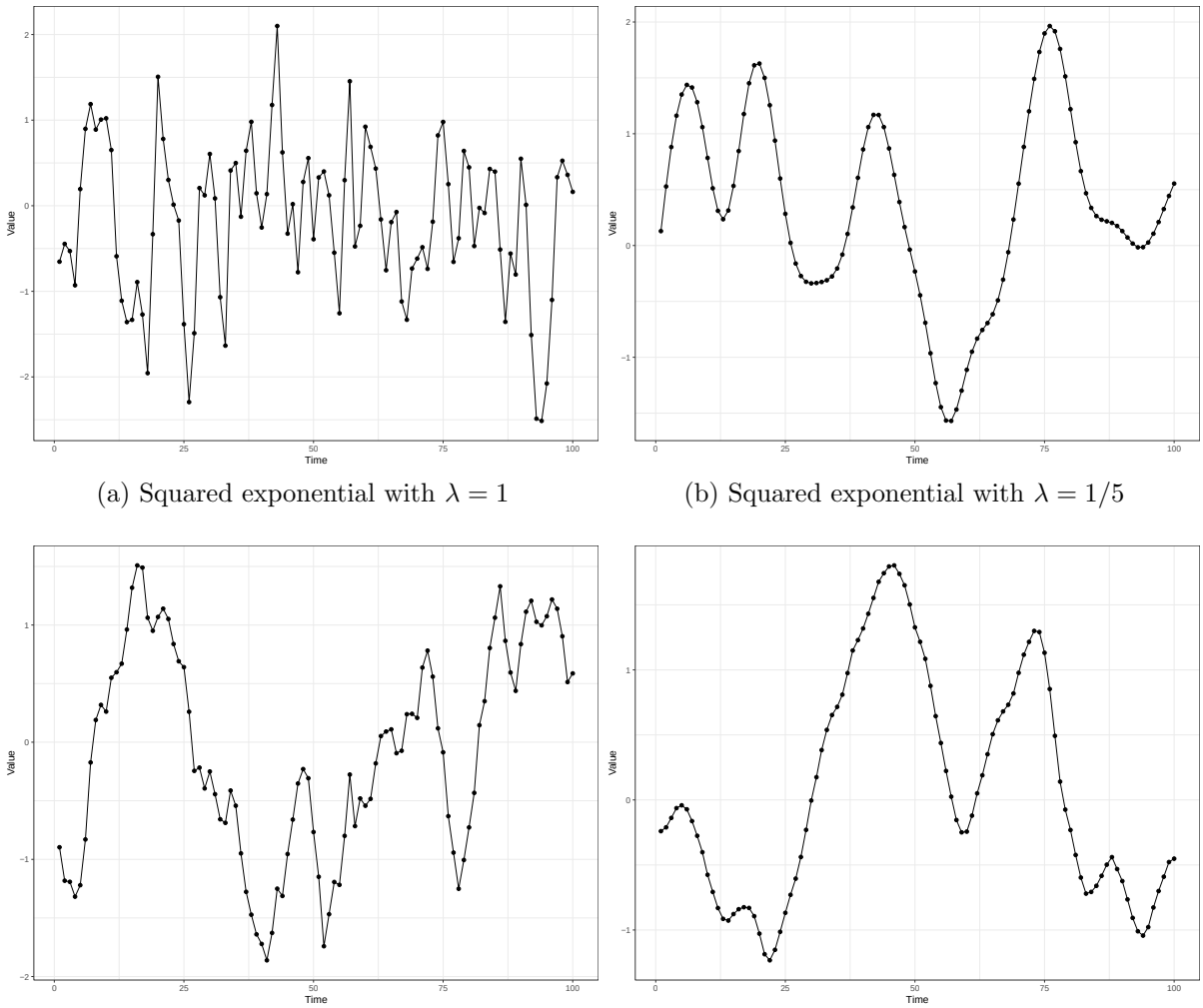
so the smoother the fit. The Matern covariance function is similar to the squared exponential function, except that it allows for further flexibility in choosing the level of differentiability. It is expressed as

$$k(t, t+s) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|s|}{\lambda} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{|s|}{\lambda} \right). \tag{5}$$

Here, Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and λ and ν are non-negative parameters of the covariance. A Gaussian process with Matern covariance has sample paths that are $\lceil \nu + \frac{1}{2} \rceil$ times differentiable.[1] The λ parameter has a similar interpretation as in the squared exponential function — the smaller the value for λ, the smoother the fit. As $\nu \to \infty$, the Matern function converges to the squared exponential covariance function. Gaussian and ARMA processes are explicitly linked. Taking $\nu = \frac{1}{2}$, results in functions that are only once differentiable, and correspond to the Ornstein–Uhlenbeck process, the continuous time equivalent of an AR(1) (Roberts et al., 2013).

[1] The $\lceil$ notation refers to the ceiling of $\nu + \frac{1}{2}$.

The varying amount of smoothness in a Gaussian Process is illustrated in Figure 2. Graphs a) and b) show Gaussian processes generated with a squared exponential covariance function. The smaller the value for λ, the smoother the process. For processes generated with a Matern covariance function (graphs c) and d)), smoothness is also influenced by the differentiability parameter, ν. This flexibility in smoothness is useful for fitting to a wide range of time series data.



(a) Squared exponential with $\lambda = 1$

(b) Squared exponential with $\lambda = 1/5$

(c) Matern with $\lambda = 1/5$ and $\nu = 1$

(d) Matern with $\lambda = 1/5$ and $\nu = 3$

Figure 2: Gaussian processes with different covariance functions and parameter values

2.5 Penalized splines regression

A spline is a piece-wise polynomial with pieces defined by a sequence of knots k. While splines have a very simple form at the local level, they offer a lot of flexibility in modeling functions in a smooth way. In particular, we consider cubic splines i.e. splines constructed of piecewise third-order polynomials, which are commonly used as a basis for regression in demographic contexts.

In splines regression, cubic basis-splines, or B-splines, are used in a regression

framework:

$$X_t = \sum_{k=1}^{K} b_k(t)\alpha_k$$

The b_ks are fixed; $b_k(t)$ is equal to the value of the kth B-spline function evaluated at time point t. The α_ks are the spline coefficients and need to be estimated.

Splines regression is illustrated in Figure 3. In the figure, each b_k is represented at a different color at the bottom of the chart. Spline placement is determined by the knot points k which are indicated by gray dotted vertical lines. Knot points occur when the spline function is at its maximum. The estimated X_t at a particular point t (as shown by the red line) is given by a linear combination of the splines b_k at time t and the estimated coefficients α_k.

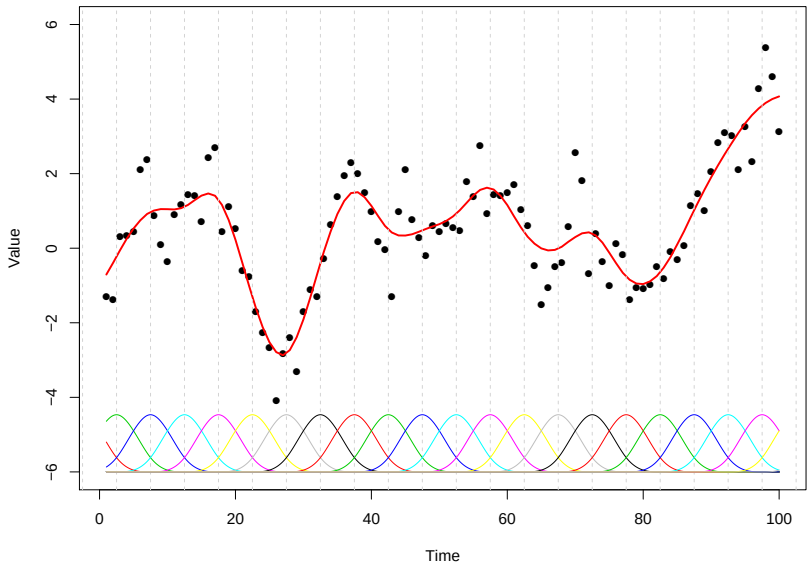


Figure 3: First-order Penalized splines regression

Splines regression models are very flexible and the smoothness of the fit can vary substantially. One way to control the smoothness of the fit is to change the spacing of the knot points; when the knots are further apart, there are fewer polynomial pieces and so the fit is more smooth. An alternative way of controlling the smoothness of the fit of X_t is to penalize differences in adjacent spline coefficients, α_k. These are called Penalized, or P-splines (Currie and Durban (2002); Eilers and Marx (1996)). In this technique, the knot spacing is taken as constant, and smoothness is imposed by the coefficient penalization. For example, the first order differences in the spline coefficients can be penalized, in which case the model set up includes the specification

$$\alpha_k \sim N(\alpha_{k-1}, \sigma_\alpha^2). \tag{6}$$

The smaller the value for the variance term, σ_α^2, the smoother the fit. Higher order differences in the coefficients can also be penalized, for example a second-order

penalization is

$$\alpha_k \sim N(2\alpha_{k-1} - \alpha_{k-1}, \sigma_\alpha^2). \tag{7}$$

For a constant knot spacing, higher-order penalizations will always produce a smoother fit.

3 Comparison of methods

The three model families are based on different assumptions and produce a smooth time series in different ways. This section highlights some of the main similarities and differences between the methods.

For each method, different parameters control the smoothness of the function (Table 1). The larger the values of ρ, θ and ν, the smoother the process. In contrast, the smaller the values for λ and σ_α, the smoother the fit. This can be seen directly from the formulas for the covariance in each particular case (Equations 2–5), as the greater the covariance between points, the smoother the fit. A smaller variance in Equations 6 and 7 for P-splines means that the differences between the spline coefficients are small.

Broadly, important distinctions between the methods occur when looking at which process is smoothed; stationarity; differentiability; and the covariance function.

3.1 Where the smoothing occurs

For ARIMA and Gaussian processes, the underlying process X_t is smoothed: assumptions are made about how different X_t points are related over time, and parameters in the model allow the process to be more or less smooth. In contrast, in P-splines regression, a line is fit to the underlying process using splines regression, and the parameters governing that line are smoothed (Table 1).

This difference has implications for how we think about characteristics of the underlying process. It is relatively intuitive to think about a time series process following an ARMA or Gaussian process: distortions being drawn from a Normal distribution, with distortions close together being more similar than those further apart. It is less intuitive to think about what the underlying process might look like in the P-splines scenario, as no explicit assumptions are made - a line is fit to a time series, and then the smoothness of that line is optimized.

3.2 Stationarity

A stationary time series is one whose statistical properties such as the mean, variance and autocorrelation are constant over time. The ARMA and Gaussian process models are stationary: there is a closed form expression for the covariance (which does not change with time t), and the processes will converge to a fixed mean and fixed variance.

In contrast, the P-splines models as defined above are not stationary. Once a P-splines regression is fitted, the empirical covariance between points in the X_t series can be calculated (Eilers et al., 2015), but there is no closed form expression for the covariance. In practice the non-stationarity has the most noticeable effect on projections of time series into the future. Consider a scenario where we are

interested in projecting forward from $t = 0$. If we assume a value for the variance of the first α_k, one can derive an expression for the covariance between α_t and α_{t-1}. For example, for first order penalization:

$$\alpha_t = \alpha_{t-1} + \varepsilon$$

with $Var(\varepsilon) = \sigma^2$ and $E(\alpha_0) = 0$. Put $Var(\alpha_0) = \sigma_0^2$. Then

$$\begin{aligned} Var(\alpha_t) &= Var(\alpha_{t-1}) + Var(\varepsilon) \\ &= Var(\alpha_{t-2}) + 2 \cdot Var(\varepsilon) \\ \cdots & \\ &= Var(\alpha_0) + t \cdot V(\varepsilon) \\ &= \sigma_0^2 + t \cdot \sigma^2 \end{aligned}$$

Then

$$\begin{aligned} Cov(\alpha_t, \alpha_{t-1}) &= Var(\alpha_{t-1}) + Cov(\alpha_{t-1}, \varepsilon) \\ &= Var(\alpha_{t-1}) \\ &= \sigma_0^2 + (t-1) \cdot \sigma^2 \end{aligned}$$

Note that the covariance expression depends on t so the process is not stationary.

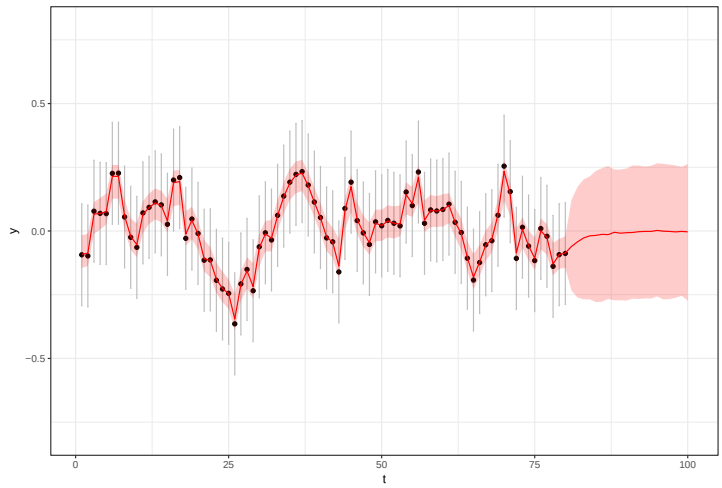
For example, Figure 4 shows the same time series of points, fitted with an AR(1) model and first-order splines regression model. The fit has been projected forward twenty periods. As a consequence of the non-stationarity of the splines process, the uncertainty intervals around the projections are much larger than for the AR(1) model, and continue to increase.

3.3 Differentiability

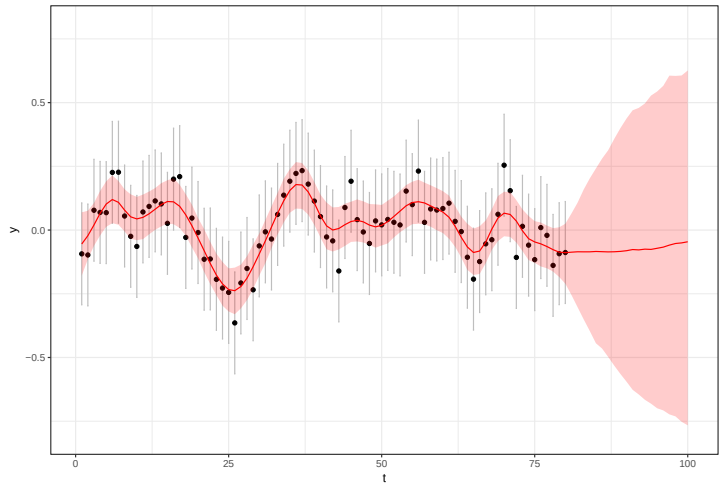
The differentiability of the underlying process affects how smooth the resulting fit appears. The higher the number of times the process is differentiable, the more the resulting fits are 'rounded'. ARMA processes are only differentiable once, which means that fits of ARMA processes can appear quite jagged (see for example, the AR(1) fit in Figure 4). The differentiability of P-splines regression models is determined by the order of the splines used; in the case of cubic splines, the process is twice differentiable. Gaussian processes with a squared exponential covariance function infinitely differentiable, while the differentiability of those with Matern covariance function can be varied by specifying different values of ν. As mentioned above, as $\nu \to \infty$, the Matern function converges to the squared exponential covariance function, and becomes infinitely differentiable.

3.4 Covariance function

A covariance function describes how X_t and X_s are related to each other, where t and s are different points in time. Intuitively, a covariance function should be of the form such that if $t \approx s$, then the covariance function approaches a maximum, meaning that X_t and X_s are very similar values. As the time points t and s get further apart, the covariance function approaches zero, and there is no dependency between the two points.



(a) AR(1) fit and projection



(b) First-order P-splines fit and projection

Figure 4: Two different smoothing functions fit on the same data. Fit 1 is an AR(1) model. Fit 2 is a first-order penalized splines regression. The fits have been projected forward twenty periods. The red line represents the mean estimate, and corresponding shaded area the 95% Bayesian credible intervals.

The covariance functions associated with all three methods have this general form. Indeed, this sort of covariance structure forms the basis of temporal smoothing - points closer to each other in time are more similar to those that are far apart. However, the covariance functions associated with each method differ in their rate of decay over time. This in turn affects the smoothness of fit. The higher the covariance between points, the smoother the fit of the X_t series.

To illustrate the relative rate of decay of the covariance functions, each method was fit to time series of the same process, and estimated parameter values were recorded. This was repeated 1,000 times, and the mean parameter estimates were calculated. Using these values, the estimated covariance between points over time can be calculated.

Figure 5 shows the correlation over time for each of the models fit to the same

time series: an ARMA(1,1) process with $\rho = 0.7$ and $\theta = 0.1$. The P-splines methods have the slowest decay, and therefore produce the smoothest fits. The correlation between points falls to approximately zero the fastest with Gaussian processes, especially with squared exponential function. While the correlation initially drops off relatively quickly for the ARMA models, after a distance of three time points, the correlation is higher than for the Gaussian process methods.

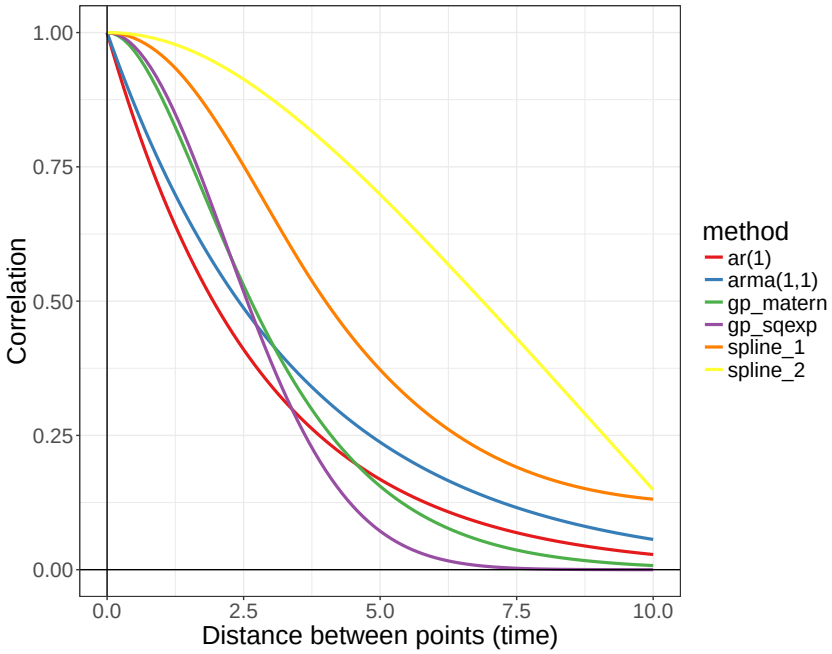


Figure 5: Correlation between points with increasing distance. Parameters for each method were based on the estimated fit on an ARMA(1,1) time series simulation with $\rho = 0.7$ and $\theta = 0.1$.

Table 1: Summary of methods

Method	Smoothing parameter(s)	What is smoothed	Stationary?	Differentiability	Covariance rate of decay
AR(1)	ρ	X_t	yes	once	medium
ARMA(1,1)	ρ, θ	X_t	yes	once	medium
GP with squared exponential	λ	X_t	yes	infinite	fast
GP with Matern	λ, ν	X_t	yes	depends on ν	fast
P-splines	σ_α	α_k	no	twice (with cubic splines)	slow

4 Comparing different data availability scenarios

The following sections aim to illustrate sensitivities in estimates, projections and uncertainty levels to different choices of temporal smoothers. The focus is on two different data scenarios which are examples of demographic estimation problems where quantifying the uncertainty around estimates and projections is important.

The first is the situation where data are readily available and considered to be good quality; however the event of interest is relatively rare and so the stochastic variability around the data are high. An example of this estimating mortality at the subnational level in developed countries. Mortality schedules in small areas are often highly erratic and may have zero death counts. In these situations, models are employed to try and estimate the underlying true mortality rates (Congdon et al. (1997); Alexander et al. (2016)).

In the second situation, there are limited data available on an indicator of interest. This is a common scenario when estimating time series for demographic and health indicators in developing countries. For example, 47% of the 193 UN-member countries have three or less observations of maternal mortality rates between the period 1990-2012.

Fitting models in the two data availability scenarios are investigated through both simulated data and two real data case studies.

4.1 The `distortr` package

The computational tools used to investigate the different methods are available through the R package accompanying this paper, `distortr`.[2] The code builds on the methods discussed above, providing tools to investigate the different behavior of methods in a range of data scenarios. The package consists of two main parts:

1. Functions to simulate time series of distortions and fit and validate models on simulated data.
2. Functions to fit Bayesian hierarchical models to datasets with observations from multiple countries. The user can specify the type of temporal smoother to fit to the data.

All models are fitted in a Bayesian framework using the statistical software R. Samples were taken from the posterior distributions of the parameters via a Markov Chain Monte Carlo (MCMC) algorithm. This was performed through the use of JAGS software (Plummer, 2003).

5 Comparison using simulated data

5.1 Setup

Methods were first fit to, and evaluated on, simulated data designed to mimic the two data scenarios discussed above. For each scenario, a time series of 100 time periods of data was simulated from each of the following processes:

[2] `https://github.com/MJAlexander/distortr`.

- AR(1) with $\rho = 0.8$;
- ARMA(1,1) with $\rho = 0.8$ and $\theta = 0.2$;
- Gaussian process with squared exponential covariance, $\lambda = 1$; and
- Gaussian process with Matern covariance $\lambda = 1/5$ and $\nu = 2$.

In each case, the last ten periods of observations were removed before fitting and the time series was projected to the full 100 periods. This was done in order to investigate differences in projections. In addition, the following alterations were made to mimic the two data availability scenarios:

1. For the small population scenario, it was assumed that standard errors around the observed data were equal to 1.
2. For the sparse data scenario, 50% of the data in the first 90 periods (i.e. 45 observations) were removed before fitting.

Each of the methods was fit to the simulated time series above (AR(1), ARMA(1,1), Gaussian process with squared exponential covariance, Gaussian process with Matern covariance, first-order P-splines and second-order P-splines). Model fits were evaluated on different metrics to understand how much it matters if the 'wrong' model is fit to a particular underlying process.

Time series of data was simulated 1000 times for each underlying process and models were fitted to each simulation. Model performance was then assessed on the average performance across all simulations.

5.2 Model comparison

Several metrics were considered to assess different aspects of model performance. The root-mean-squared-error (RMSE) was calculated to assess the average error between fitted and true values:

$$RMSE = \sqrt{\sum_{t=1}^{T} \frac{(\hat{X}_t - X_t^*)^2}{T}},$$

where $\hat{X}_t$ is the estimated value at time t from the model of interest and X_t^* is the true value at time t and T is the total number of observations.

In addition, the sharpness of the uncertainty intervals was measured by:

$$n_t = \sum_{t=1}^{T} \frac{(r_t - l_t)}{T}$$

where l_t and r_t are the lower and upper bounds of a 95% Bayesian credible interval, respectively. The average width was measured separately for estimated and projected values.

5.3 Results

Tables 2 and 3 show the RMSE for each method (row) fit to different underlying processes for data scenario 1 (small populations) and 2 (sparse data), respectively.

Table 2: RMSE, Scenario 1 (small populations)

	Process			
Method	AR(1)	ARMA(1,1)	GP sq exp	GP Matern
AR(1)	0.35	0.47	0.66	0.07
ARMA(1,1)	0.36	0.48	0.63	0.06
GP sq exp	0.40	0.50	0.57	0.04
GP Matern	0.64	0.80	0.68	0.06
Splines 1	0.48	0.64	0.64	0.05
Splines 2	0.60	0.75	0.60	0.05

Table 3: RMSE, Scenario 2 (sparse data)

	Process			
Method	AR(1)	ARMA(1,1)	GP sq exp	GP Matern
AR(1)	0.46	0.57	0.28	0.01
ARMA(1,1)	0.47	0.60	0.25	0.01
GP sq exp	0.99	1.52	0.58	0.01
GP Matern	1.29	1.51	0.66	0.01
Splines 1	0.62	0.71	0.24	0.01
Splines 2	0.91	1.21	0.52	0.01

In general, processes are best described by the model from the same family. The P-splines models produce relatively smooth fits and generally have the highest RMSEs. Model misfit is more prevalent in Scenario 2, where there are incomplete time series.

Tables 4 and 5 show the average width of the 95% uncertainty intervals within the estimates and projections. Several observations can be made. Firstly, for the estimated time series points, the ARMA processes have the widest uncertainty intervals, followed by Gaussian processes and then P-splines. In contrast, when it comes to projection, the order is switched: in general ARMA processes have the smallest uncertainty interval width, while P-splines have the largest. Differences across methods are more noticeable in scenario 2, where the data are sparse.

Table 4: Average uncertainty interval width, Scenario 1

Method	Estimates	Projections
AR(1)	1.73	3.14
ARMA(1,1)	1.69	3.38
GP sq exp	1.64	3.50
GP Matern	1.51	3.60
Splines 1	1.46	4.74
Splines 2	1.42	7.91

These differences in the width of uncertainty intervals is related to both the shape of the covariance function for different methods and the stationarity of the process. As seen in Figure 5, the covariance function for the ARMA models decreases quickly relative to the other methods. This means that, in general, there is less assumed correlation between points through time, and so there are more possi-

Table 5: Average uncertainty interval width, Scenario 2

Method	Estimates	Projections
AR(1)	1.83	4.27
ARMA(1,1)	1.81	4.37
GP sq exp	1.82	4.33
GP Matern	1.78	4.99
Splines 1	1.06	13.3
Splines 2	1.04	24.29

ble paths for estimates to follow. This makes the uncertainty around the estimates higher.

In terms of projections, the P-splines processes are not stationary, which means uncertainty around the projections continues to increase over time. Note that in the simulation setup the time series were projected forward ten periods. The observation that uncertainty is higher around P-splines projections does not necessarily hold for shorter-term projections. Indeed, shorter-term projections for P-splines may have smaller amounts of uncertainty due to the higher implied covariance between points.

6 Comparison using case studies

The two different data scenarios were also investigated through fitting models to real data. The example for Scenario 1 (small populations) was Australian regional mortality, while the example for sparse data availability was global estimation of antenatal care.

Each of the models described above were fit to data in the two case studies. The relative performance of the models was assessed using the deviance information criterion (DIC). This criterion measures a combination of model fit and a penalization based on the number of parameters. Also considered was the root-mean squared relative difference between each pair of models:

$$RMSRD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{X}_{tk} - \hat{X}_{tl})^2}{\bar{X}_{tk}}},$$

where $\hat{X}_{tk}$ and $\hat{X}_{tl}$ are the estimated values at time t for models k and l. The average width of uncertainty intervals was also measured as above.

6.1 Small populations: Australian regional mortality

The dataset used to explore fitting to small populations was mortality rates by Local Government Area (LGA) in the state of New South Wales (NSW) in Australia. Data were obtained from the Australian Bureau of Statistics (ABS, 2016). There are 157 LGAs in NSW, which is the most populous state in Australia. The population size varies widely across the LGAs, ranging from less than 2,000 people in some rural areas, to almost 340,000 people in Blacktown (an area in Western Sydney).

Standard errors around the observed death rates were obtained by assuming deaths follow a Poisson distribution. The mortality rate was estimated for all areas

over the period 2000-2020. Data were available each year between 2000 and 2015 and the last five years were projected forward. Each method was fit within a Bayesian hierarchical framework, using functions as part of the `distortr` package. The hierarchical setup allows information about smoothness and autocorrelation to be pooled across areas.

6.1.1 Results

Results of fitting to the NSW mortality dataset suggest that model performance is relatively similar across all methods (Tables 6, 7 and 8). The DIC (Table 6) is the lowest for the AR(1) model, suggesting that this is the preferred model in this data scenario. Note that the AR(1) and a Gaussian process with a squared exponential models are the most parsimonious, as only one parameter needs to be estimated per region (ρ and λ, respectively). In contrast, P-splines regression requires the estimation of the smoothing parameter (σ_α) and also as many spline coefficients α_k as there are knots. The DIC values suggest that any improvement in fit with P-splines was not enough to offset the increase in parameters that need to be estimated.

Table 6: DIC for each method, Scenario 1

Method	DIC
AR(1)	28,084.27
ARMA(1,1)	28,781.48
GP sq exp	28,627.09
GP Matern	28,566.73
Splines 1	28,812.79
Splines 2	28,706.46

The root-mean-squared differences between methods (Table 7) also suggest relatively small differences in the point estimates across methods. The differences were the smallest within the modeling families; i.e. AR(1) and ARMA(1,1) were most alike, the same being for the Gaussian process and P-spline pairs. The largest differences were between the ARMA methods and second-order P-splines.

Table 7: Root-mean-squared differences between methods, Scenario 1

	AR(1)	ARMA(1,1)	Splines 1	Splines 2	GP Matern
AR(1)	-	-	-	-	-
ARMA(1,1)	0.0070	-	-	-	-
Splines 1	0.0334	0.0331	-	-	-
Splines 2	0.0403	0.0403	0.0175	-	-
GP Matern	0.0311	0.0313	0.0141	0.0257	-
GP sq exp	0.0322	0.0321	0.0146	0.0281	0.0051

Comparing the average width of uncertainty intervals around the projections (i.e. 2016-2020) again suggests that all the methods were fairly similar. While the ARMA models have the largest average uncertainty, the second order P-splines has the widest uncertainty interval around the last year. Again, this is related to the

interactions between the covariance function and stationarity. The rate of decay of the ARMA covariance functions is relatively fast, so the uncertainty around projections also increases relatively fast; however, a stationary variance is reached. In contrast, uncertainty in P-splines projections continues to increase over time.

Table 8: Average width of projection 95% uncertainty intervals, Scenario 1

Method	Average interval width	Last interval width
AR(1)	0.159	0.182
ARMA(1,1)	0.166	0.192
Splines 1	0.123	0.183
Splines 2	0.137	0.210
GP Matern	0.117	0.172
GP sq exp	0.103	0.188

The graphs shown in Figure 6 illustrate how each of the methods fit the NSW mortality data. The ARMA methods are less smooth than other methods due to being lower-order differentiable process. Uncertainty around estimates in the period 2010-2015 is fairly similar, but lowest for the P-spline methods.

6.2 Sparse data: global data on antenatal care visits

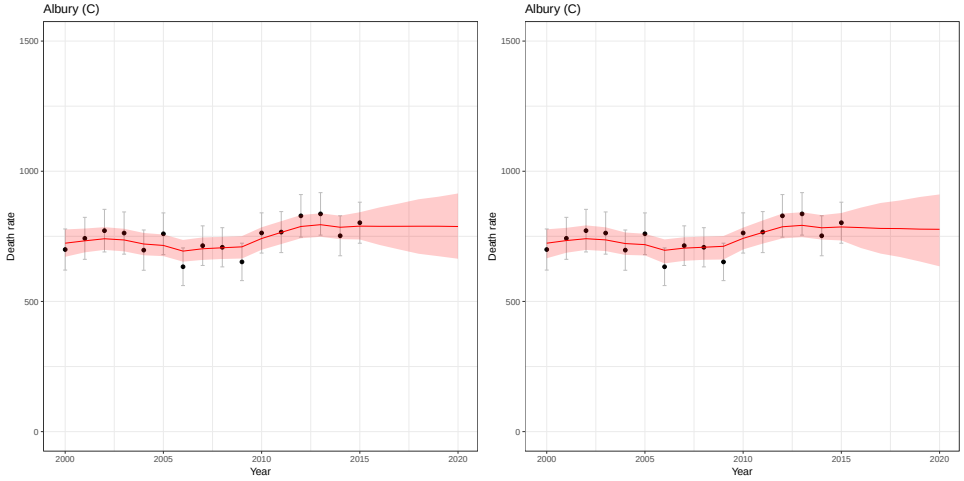
The second case study explored fitting to sparse data. The example dataset is a global database on antenatal care. Specifically, the indicator of interest is the percentage of women aged 15-49 years attended at least four times during pregnancy by any provider (hereafter referred to as ANC4). The database was provided by the World Health Organization (WHO, 2017) and was compiled from publicly available data. The aim of modeling is to produce estimates of ANC4 (and uncertainty) for all countries from over the period that corresponds to the first observation year in that country to the year 2015.

Data are available for 150 countries, with the number of observations and observation time period varying substantially by country. Many countries only have very limited data, with only one or two observations.

Modeling was done on the logit scale to ensure estimates produced are between zero and one. Standard errors were obtained assuming the proportion of women seeking antenatal care follows a binomial distribution. Given many of the data sources may not be representative of the broader population, it is likely that non-sampling errors also need to be accounted for. For the ANC4 data, there are three source types: survey, administrative, and other. The majority of the data come from surveys or administrative sources, but there are also some data from other sources such as the Pan American Health Organization, Health Information and Analysis Project. Non-sampling error is estimated in the modeling process. In practice, administrative data has the smallest non-sampling error, followed by surveys and then other sources.

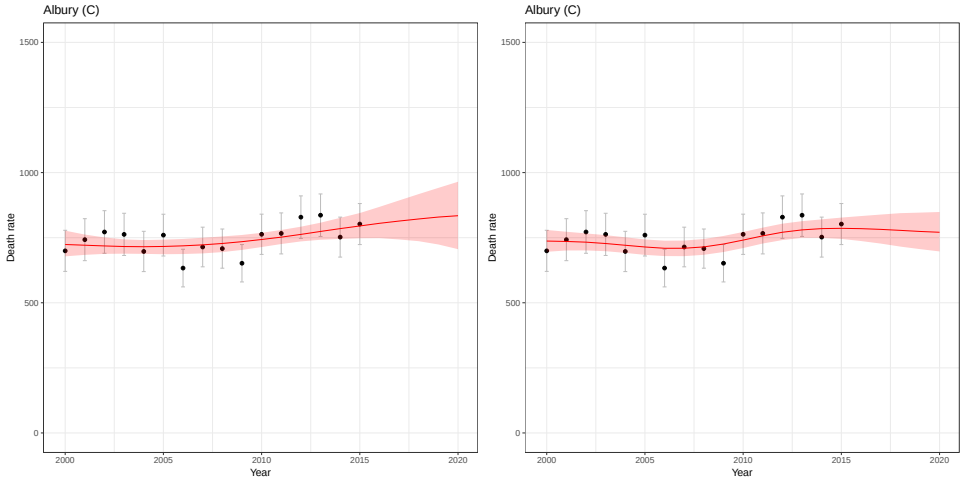
6.2.1 Results

In contrast to Scenario 1, the DIC for different methods varies quite substantially (Table 9). The Gaussian process methods and second-order P-splines have the best DIC results, while the ARMA models perform the worst.



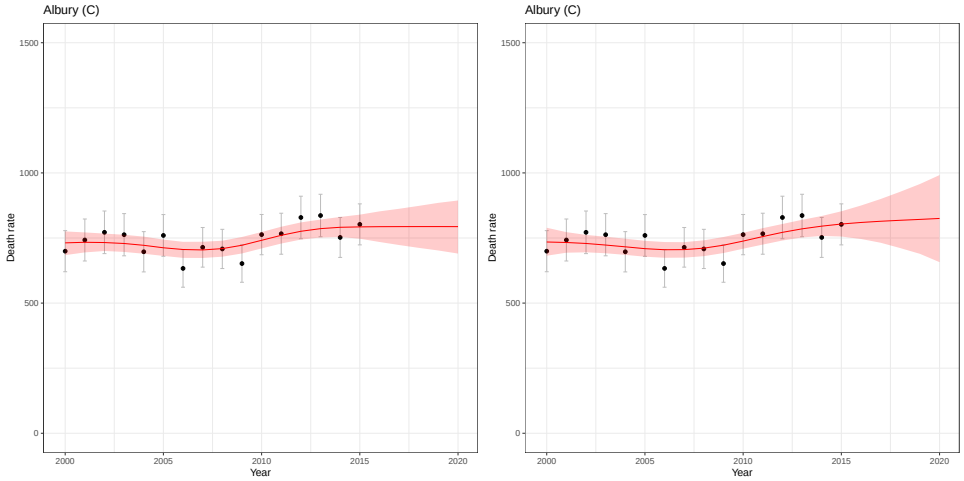
(a) AR(1)

(b) ARMA(1,1)



(c) GP, squared exponential

(d) GP, Matern



(e) First-order P-splines

(f) Second-order P-splines

Figure 6: Estimates and projections for the mortality rate in Albury

Table 9: DIC for each method, Scenario 2

Method	DIC
AR(1)	5,630.893
ARMA(1,1)	7,009.201
GP sq exp	1,711.868
GP Matern	1,832.049
Splines 1	3,299.185
Splines 2	1,753.965

Comparing the root-mean square differences suggests that on average the point estimates across the different methods vary more widely than in Scenario 1, but are still are fairly similar (Table 10). Again, the pairs of models within the same family are the most similar.

Table 10: Root-mean-squared differences between methods, Scenario 2

	AR(1)	ARMA(1,1)	Splines 1	Splines 2	GP Matern
AR(1)	-	-	-	-	-
ARMA(1,1)	0.0212	-	-	-	-
Splines 1	0.0298	0.0245	-	-	-
Splines 2	0.0360	0.0336	0.0258	-	-
GP Matern	0.0320	0.0293	0.0149	0.0270	-
GP sq exp	0.0320	0.0297	0.0151	0.0265	0.044

Due to data sparsity and the presence of non-sampling errors, the average uncertainty around estimates and projections is much higher in the ANC4 case. In addition, the variation in uncertainty across methods is much greater (Table 11). Average uncertainty for ARMA methods is around 1.5 times more than the P-splines methods, and this observation also holds for the projections.

Table 11: Average width of projection 95% uncertainty intervals, Scenario 2

Method	Average interval width
AR(1)	0.462
ARMA(1,1)	0.494
Splines 1	0.399
Splines 2	0.287
GP Matern	0.354
GP sq exp	0.343

Figure 7 illustrates the differences across methods fit to the ANC4 data for Bolivia. There are multiple data sources; data in earlier periods are from surveys, while the last two observations are from another, non-administrative source (Pan American Health Organization, Health Information and Analysis Project). As mentioned above, the estimated non-sampling error associated with survey data is smaller than other sources, and so the overall fit is less influenced by the last two observations.

In general, there are two main differences across the methods. Again the the smoothness of fit varies, which is influenced by the differentiability of the process. For example, the Gaussian process with squared exponential process is infinitely differentiable, which leads to a very smooth fit. In contrast, AR(1) and ARMA(1,1) fits can have more sharp points, and this is reflected in the shape of the uncertainty bands. In addition, methods with covariance functions with relatively slow rates of decay have smoother fits, and also leads to uncertainty intervals being generally smaller and smoother. This is particularly the case for the second-order P-splines, whose uncertainty interval does not even include the last two data points.

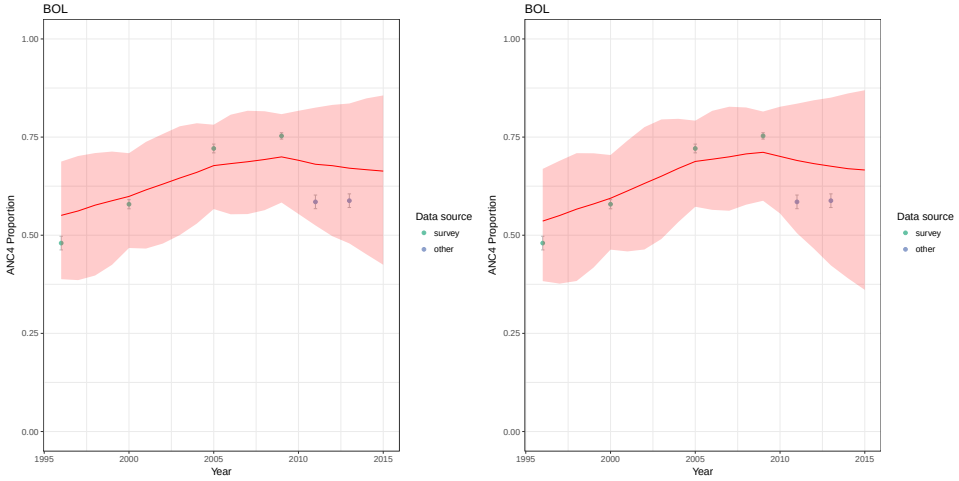
7 Discussion

The development of methods to estimate and project demographic and health indicators is important to help monitor and understand trends and inequalities over time. In addition to classical covariate-based regression models, it is often the case that estimation methods require a temporal component which is flexible enough to capture data-driven non-linearities in trends over time. These temporal methods are also important in capturing uncertainties in the inherent stochastic process that is underlying the data. This paper presented a review of three families of temporal smoothing methods commonly found in demographic literature. After highlighting similarities and differences in the underlying assumptions of each model family, the sensitivity of fits to model choice were exploring using both simulated and real data.

One of the most important observations was that model choice affects not only point estimates but also affects — and perhaps even to a larger extent — the implied uncertainty around estimates and projections. In general, simple lower-order ARMA processes such as AR(1) and ARMA(1,1) appear to produce larger uncertainty bands around estimates and short-term projections. On the other hand, P-splines regression techniques provide relatively small uncertainty bands around estimates, but the uncertainty around longer-term projections quickly increases more than any of the other methods considered. Differences are particularly noticeable in data-sparse scenarios.

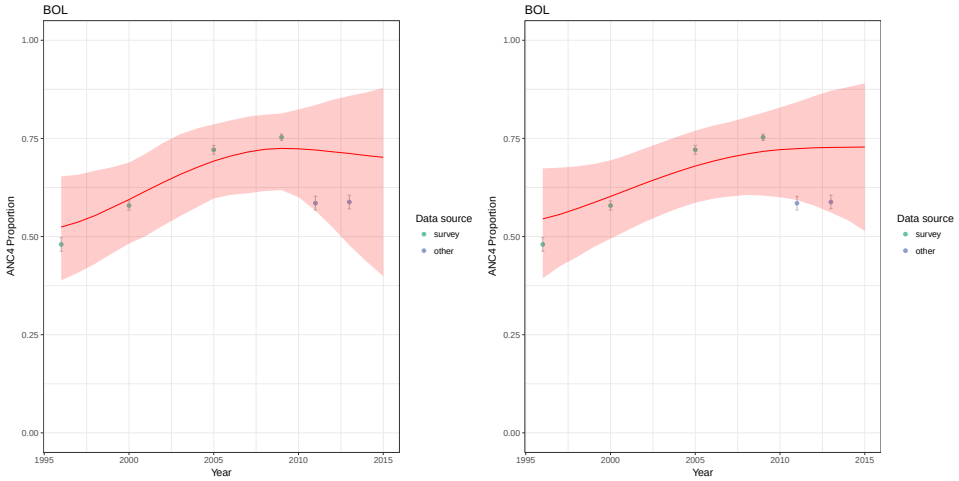
While demographic modeling has traditionally been based on deterministic methods, there has been an increasing move towards including stochastic methods to fully capture uncertainty in estimates and projections (Girosi and King (2008); Raftery et al. (2012); UNPD (2017)). Indeed, producing and reporting estimates of uncertainty is important in demographic research because it communicates a measure of level of confidence we have about trends in the past and what is likely to happen in the future. However, this work highlights the importance of considering different sources of uncertainty and what types are included in any estimation. We need to think about difference sources of error and uncertainty in data and also model choice. Just because uncertainty levels are estimated, does not mean that they are the 'right' levels, and could be driven by model choice, which often is relatively arbitrary.

In terms of model choice, there are many factors that affect what is appropriate to use in different data scenarios. This paper has focused on the temporal smoothing aspect only, but of course important modeling decisions must also be made with regards to any covariate-based modeling framework. In light of the analysis presented in this paper, it may be useful to consider the following to help inform



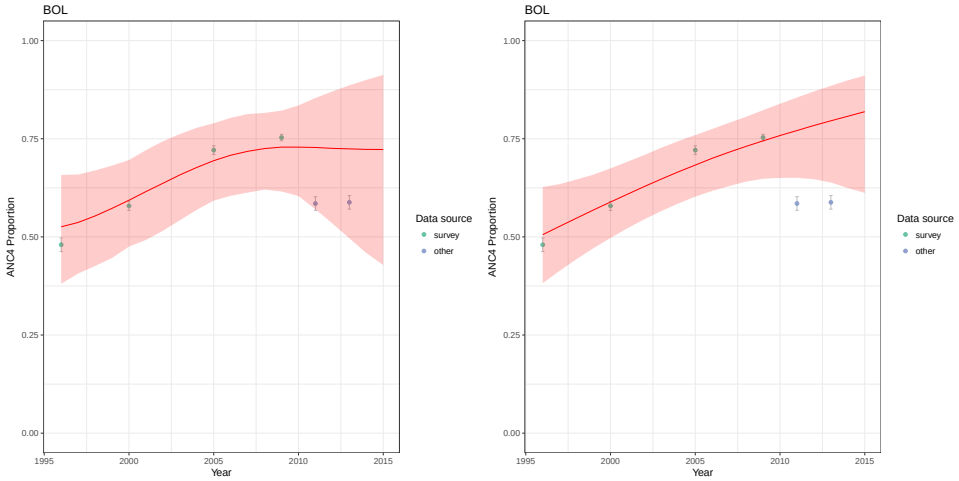
(a) AR(1)

(b) ARMA(1,1)



(c) GP, squared exponential

(d) GP, Matern



(e) First-order P-splines

(f) Second-order P-splines

Figure 7: Estimates and projections for ANC4 in Bolivia

the choice of temporal smoothing method, and motivate future research.

An important consideration is whether there is prior information or beliefs about the data generating process underlying the outcome of interest. For example, if data on the outcome of interest is readily available, it may be feasible to explore the autocorrelation structure implied by the data, through examination of the sample autocorrelation and sample partial autocorrelation functions. Results from this analysis could be used to inform model choice. If the dataset of interest contains countries or regions that have different data availability, then patterns in the high-quality data countries could be used to inform other countries in a hierarchical setup.

If the available data are sparse for most population of interest, then it may not be possible to get a good idea of the autocorrelation structure. Additionally, as the analysis in this paper highlighted, situations with sparse data are the most sensitive to model choice. One modeling approach could be to assess the performance of a range of methods using both in- and out-of-sample metrics; for example, testing on a validation dataset to assess error and coverage of uncertainty intervals (see for example Alkema and New (2014)). If there is a stand-out method then that could be used; otherwise, the most parsimonious model may be the most appropriate.

There have been large advancements in methods used in demographic estimation which allow for more robust estimates to be produced, giving an indication of how certain we are about these estimates. It is important to be aware of the many sources of uncertainty that occur when choosing a modeling framework, and how sensitive the outcomes may be to these choices. Greater transparency in why modeling decisions are made is important in improving the understandability and reproducibility of demographic research.

References

ABS (2016). 3302.0 - deaths, australia, 2016. Available at: `http://www.abs.gov.au/ausstats/abs@.nsf/mf/3302.0`.

Alexander, M. and L. Alkema (2016, December). Global Estimation of Neonatal Mortality using a Bayesian Hierarchical Splines Regression Model. *ArXiv e-prints*.

Alexander, M., E. Zagheni, and M. Barbieri (2016, July). A Flexible Bayesian Model for Estimating Subnational Mortality. *ArXiv e-prints*.

Alkema, L., D. Chou, D. Hogan, S. Zhang, A.-B. Moller, A. Gemmill, D. M. Fat, T. Boerma, M. Temmerman, C. Mathers, and L. Say (2016, 2017/07/27). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un maternal mortality estimation inter-agency group. *The Lancet 387*(10017), 462–474.

Alkema, L., V. Kantorova, C. Menozzi, and A. Biddlecom (2013, 2017/07/27). National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. *The Lancet 381*(9878), 1642–1652.

Alkema, L. and J. R. New (2014). Global estimation of child mortality using a bayesian b-spline bias-reduction model. *The Annals of Applied Statistics 8*(4), 2122–2149.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Congdon, P., S. Shouls, and S. Curtis (1997). A multi-level perspective on small-area health and mortality: a case study of england and wales. *Population, Space and Place 3*(3), 243–263.

Currie, I. D. and M. Durban (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling 2*(4), 333–349.

Currie, I. D., M. Durban, and P. H. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical Modelling 4*(4), 279–298.

Eilers, P. H., B. D. Marx, and M. Durbán (2015). Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions 39*(2), 149–186.

Eilers, P. H. C. and B. D. Marx (1996, 05). Flexible smoothing with b-splines and penalties. *Statist. Sci. 11*(2), 89–121.

Finucane, M., C. Paciorek, G. Danaei, and M. Ezzati (2014). Bayesian estimation of population-level trends in measures of health status. *STATISTICAL SCIENCE 29*, 18–25.

Foreman, K. J., R. Lozano, A. D. Lopez, and C. J. Murray (2012, Jan). Modeling causes of death: an integrated approach using codem. *Population Health Metrics 10*(1), 1.

Girosi, F. and G. King (2008). *Demographic forecasting*. Princeton University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Raftery, A. E., N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences 109*(35), 13915–13921.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain (2013). Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A 371*(1984), 20110550.

UNPD (2017). World population prospects: The 2017 edition. Available at: `http://esa.un.org/wpp/`.

WHO (2017). Antenatal care (at least 4 visits). Available at: `http://www.who.int/gho/urban_health/services/antenatal_care/en/`.