Comparing Temporal Smoothers for Use in Demographic Estimation

Monica Alexander
University of California, Berkeley

IUSSP 2017
Session 143: Advances in mortality modeling
November 1, 2017

Motivation

- Need accurate estimates and projections of demographic and health indicators over time
- Important to monitor progress in health outcomes, informing policy
- In some cases trends may be unclear, because of missing or messy data
- Need to use statistical models to estimate and smooth data over time

Motivation

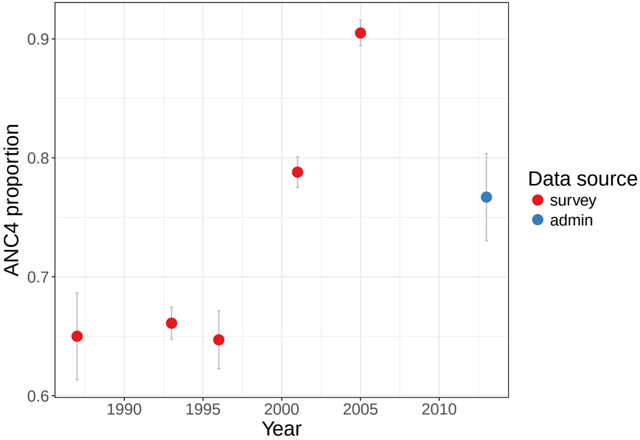


Figure: Proportion of women of reproductive age with adequate antenatal care, Paraguay

Motivation

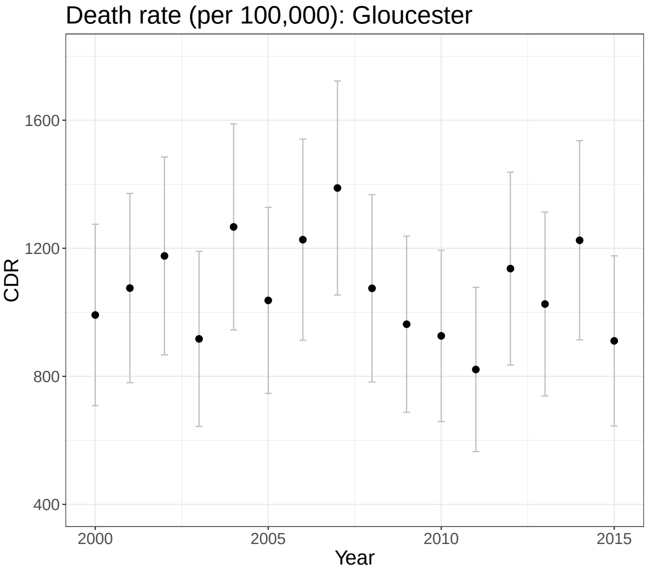


Figure: Death rates for Gloucester, New South Wales, Australia

Motivation

How to model a demographic time series? Many models in the literature have the same general format. Define θ_t to be the quantity of interest at time t. Modeled as:

$$\theta_t = f(X_t) + Z_t + \varepsilon_t$$

- Regression framework $f(X_t)$, a function of covariates X_t
- distortions Z_t capture data-driven non-linear trends over time, not otherwise captured in $f(X_t)$

Models for Z_t smooth data temporally.

Examples

- Neonatal mortality rates (Alexander et al. 2016): $f(X_t) = f(\text{U5MR})$; Z_t is modeled through P-splines regression
- Maternal mortality rates (Alkema et al. 2014): $f(X_t)$ includes GDP, skilled attendants at birth; Z_t is modeled using an ARMA(1,1) process
- Adult mortality (IHME): $f(X_t)$ includes income, education; Z_t is modeled using Gaussian process regression

While $f(X_t)$ is often justified, the choice for Z_t seems more arbitrary. But Z_t is important:

- Flexibly model data-driven trends
- Incorporate uncertainty
- Define a temporal process that can be projected forward in time

Aims

Focus on modeling the distortions, Z_t. Compare three main families that commonly occur in demographic literature:

1. ARMA models (MMR, contraceptive prevalence)
2. Gaussian Process regression (cause-specific mortality)
3. Penalized splines regression (child and adult mortality)

Aims:

- Compare theoretical differences
- Evaluate model performance and sensitivities on both simulated and real data

ARMA models

Autoregressive moving average (ARMA) models

- The autoregressive (AR) part assumes that Z_t 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.

A first-order Autoregressive process, or AR(1):

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

First-order Autoregressive Moving Average models, i.e. ARMA(1,1):

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

Gaussian Process regression

- Gaussian processes (GPs) extend multivariate Gaussian (Normal) distributions to infinite dimensionality.
- GPs can form the basis of a regression to estimate and predict new data points.

For any sequence of times, $\boldsymbol{t} = t_1, t_2, \ldots, t_n$ a GP is

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

with mean function $m(\boldsymbol{t})$ and covariance function $k(\boldsymbol{t}, \boldsymbol{t})$ (focus on squared exponential and Matern covariance functions)

Penalized splines regression

Basis-splines, or B-splines, are used in a regression framework:

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

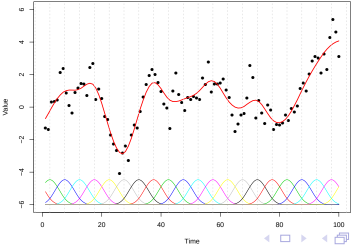
- $b_k(t)$ is equal to the value of the kth B-spline function evaluated at time point t.

Control the smoothness of Z_t by penalizing differences in adjacent spline coefficients, α_k. First-order:

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

Second-order penalization is

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



Comparison of methods

- Compare theoretically
- Compare fits, focusing on two data scenarios:
  1. High variability
  2. Limited data

Comparison of methods

- Simulation setup:
  1. Simulate time series based on process (ARMA, GP). Mimic real-world data scenarios:
     - High variability: simulate with large stochastic variance
     - Limited data: remove observations
  2. Fit different models (ARMA, GP, P-splines) to each simulation, repeat
  3. Evaluate fit (root mean square error) and width of uncertainty intervals
- Real data scenarios
  - High variability: regional mortality in Australia
  - Limited data: antenatal care in countries worldwide

Theoretical differences

Highlighting two theoretical differences and show when these matter the most. Two important differences in:

1. Stationarity
2. Covariance function

These differences affect both the **point estimates** and **uncertainty around estimates**.

Stationarity

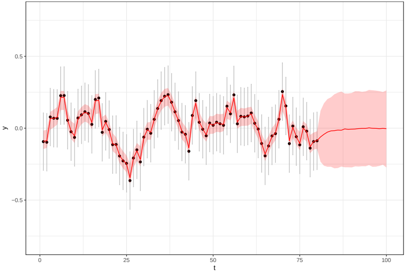
A stationary time series has constant mean, variance and autocorrelation over time.

- ARMA and GP models are fitted as stationary processes
- P-splines models as defined above are not stationary

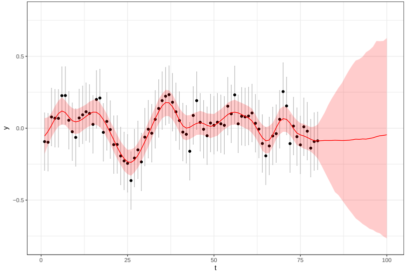
In practice stationarity has the most noticeable effect on projections of time series.

- For ARMA and GP, mean and width of uncertainty eventual converge to stable levels
- P-splines projections do not converge to a constant mean and uncertainty intervals increase

Stationarity



(b) AR(1) fit and projection



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

Figure: Two different smoothing functions fit on the same data. The fits have been projected forward twenty periods. The red line represents the mean estimate, and corresponding shaded area the 95% Bayesian credible intervals.

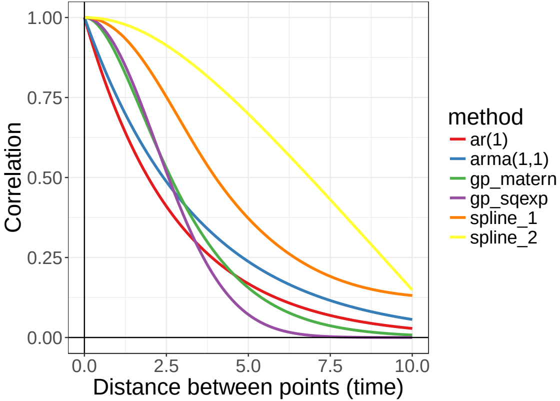
Covariance function

A covariance function describes how values at two different times are related to each other.

- Each method has a different implied covariance function
- Higher covariance between points leads to a smoother fit and smaller uncertainty intervals.

Covariance function

Figure: Correlation between points with increasing distance. Based on the estimated fit on an ARMA(1,1) time series simulation with $\rho = 0.7$ and $\theta = 0.1$.



- ARMA and GP have relatively less smooth fits and wider uncertainty intervals compared to P-splines

Covariance function

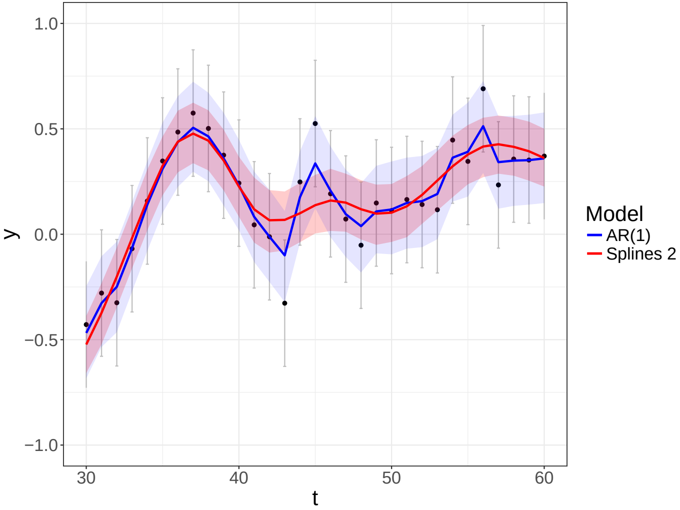


Figure: AR(1) and Second-order P-Splines

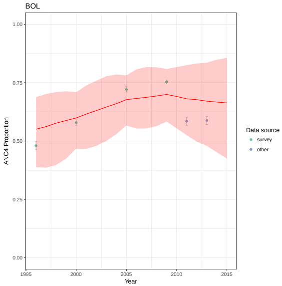
Comparison

When do these differences matter? Differences in point estimates and uncertainty are largest in limited data situations.

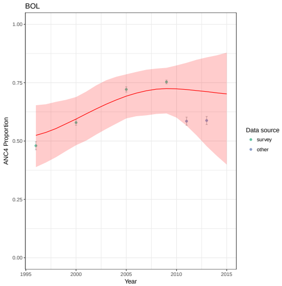
Table: Average RMSE (standard deviation) across fits using AR(1), ARMA(1,1), GP sq exp, GP Matern, 1st and 2nd-order P-Splines

Process	**Highly variable data**	**Limited data**
AR(1)	0.47 (0.12)	0.79 (0.33)
ARMA(1,1)	0.60 (0.15)	1.02 (0.45)
GP sq exp	0.63 (0.14)	0.65 (0.19)
GP Matern	0.05 (0.01)	0.01 (0.01)

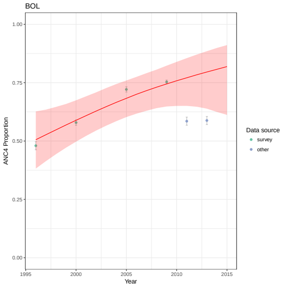
Example: limited data



(a) AR(1)



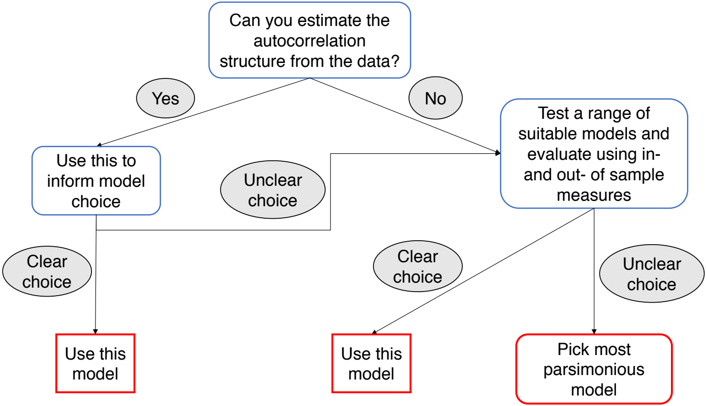
(b) GP, squared exponential



(c) Second-order P-splines

Figure: Estimates of proportion of women of reproductive age with adequate antenatal care, Bolivia

Some thoughts on modeling decisions



Summary

- Investigated three methods for temporal smoothing that are common in demographic literature
- This is initial work part of a broader project to help demographers and policymakers make informed decisions about demographic modeling
- Showed underlying theoretical differences lead to differences in both estimates and uncertainty
- Important to consider different sources of uncertainty and how estimates are generated

Acknowledgements

This work is part of a project supported by the Department of Health Statistics and Information Systems at the World Health Organization.

R package to simulate and fit models available here:
`https://github.com/MJAlexander/distortr`

Thanks!