Journal of Survey Statistics and Methodology (2023) **00**, *1–25*

MODELING PUBLIC OPINION OVER TIME: A SIMULATION STUDY OF LATENT TREND MODELS

MARTA KOŁCZYŃSKA *
PAUL-CHRISTIAN BÜRKNER

With the growing availability of multi-wave surveys, social scientists are turning to latent trend models to examine changes in social and political attitudes. Aiming to facilitate this research, we propose a framework for estimating trends in public opinion consisting of three components: the measurement model that links the observed survey responses to the latent attitude, the latent trend model that estimates a trajectory based on aggregated individual latent scores, and representativeness adjustments. We use individual-level item response theory models as the measurement model that is tailored to analyzing public opinion based on pooled data from multi-wave surveys. The main part of our analysis focuses on the second component of our framework, the latent trend models, and compares four approaches: thin-plate splines, Gaussian processes, random walk (RW) models, and autoregressive (AR) models. We examine the ability of these models to recover latent trends with simulated data that vary the shape of the true trend, model complexity, and data availability. Overall, under the conditions of our simulation study, we find that all four latent trend models perform well. We find two main performance differences: the relatively higher squared errors of AR and RW models, and the under-coverage of posterior intervals in high-frequency

MARTA KOŁCZYŃSKA is an Assistant Professor with the Institute of Political Studies of the Polish Academy of Sciences, ul. Polna 18/20, 00-625 Warsaw, Poland. PAUL-CHRISTIAN BÜRKNER is an Independent Junior Research Group Leader with the Cluster of Excellence SimTech, University of Stuttgart Universitätsstraße 32, 70569 Stuttgart, Germany.
We thank Aki Vehtari for discussions of the simulation design. We also acknowledge the computational resources provided by the Aalto Science-IT project and by the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, under computational allocation GB84-31.
This work was supported by the Polish National Agency for Academic Exchange (PPN/BEK/2019/1/00133), the Polish National Science Centre (2019/32/C/HS6/00421), and the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2075-390740016).
*Address correspondence to Marta Kołczyńska, Institute of Political Studies of the Polish Academy of Sciences, ul. Polna 18/20, 00-625 Warsaw, Poland; E-mail: mkolczynska@isppan.waw.pl.

https://doi.org/10.1093/jssam/smad024

low-amplitude trends with thin-plate splines. For all models and across all scenarios, performance improves with increased data availability, which emphasizes the need of supplying sufficient data for accurate estimation of latent trends. To further illustrate the differences between the four latent trend models, we present a case study with an analysis of trends in political trust in Hungary, Poland, and Spain between 1995 and 2019. We note the relatively weaker performance of splines compared to other models in this application and conclude by discussing factors to consider when choosing the latent trend model, and further opportunities in this line of research.

KEYWORDS: Gaussian processes; Latent trend models; Public opinion; Random walk models; Survey data; Thin-plate splines.

Statement of Significance

Researchers increasingly pool data from multiple surveys conducted over many years to estimate trajectories of mass public opinion. To systematize this research, we propose to distinguish three components of the modeling procedure: the measurement model, the latent trend model, and representativeness adjustments. We focus on latent trend models, which account for the temporal sequence of the data and estimate a trend from the aggregated individual scores. We compare four approaches: thin-plate splines, Gaussian processes, random walk models, and autoregressive models, via a simulation study and a case study with survey data on political trust from three countries. We find that the performance of the analyzed approaches depends on the shape of the true trend and data availability.

1. INTRODUCTION

Trajectories of public opinion are of interest to social science researchers as a tool for monitoring social change, and as hypothesized causes or consequences of political, social, and economic developments. National and cross-national surveys constitute a rich data source describing attitudes, values, and preferences from multiple countries over many decades. Combining these data into country time series of aggregate public opinion creates new research opportunities, yet requires a number of methodological choices that the literature has thus far not systematically discussed. To guide researchers through these choices, we propose a framework for analyzing public opinion over time by pooling existing survey datasets. The framework includes three main components: the measurement model to link the observed survey responses to the latent trait, in this case an opinion or attitude, the latent trend model that estimates a trajectory based on aggregated individual latent scores, and representativeness adjustments to account for the possible deviations from representativeness of the survey sample. Distinguishing these three components introduces structure to methodological research on the relative merits of different approaches within components, as well as on the interplays of approaches across components.

The present article focuses on the statistical validation of the middle component of the framework, namely the model for the latent trend. Prior studies proposed solutions primarily relying on random walk (RW) models (Caughey and Warshaw 2015; Claassen 2019; Solt 2020). The paper by Kołczyńska et al. (2020) used thin-plate splines (TPS). Gaussian processes (GPs), and autoregressive (AR) models constitute additional alternatives that—to the best of our knowledge—have not yet been applied or studied in this context. Our analysis constitutes the first systematic comparison of these four approaches and unifies previously disconnected model classes that researchers have proposed.

We compare the performance of models with the latent trend modeled as TPS, GP, RW, or AR process, in scenarios that differ with regard to the shape of the true latent trend, presence or absence of between-person variation, and data availability. Under the conditions of our simulation study, all four models performed well in all scenarios, even in the most challenging condition that combines small fluctuations of the true trend with data sparsity. Our results point to relatively weaker performance of AR and RW models in terms of root mean squared error (RMSE), posterior standard deviation, and absolute bias, as well as to substantial under-coverage of posterior credible intervals implied by the TPS models. GP models emerge as the option with reliably high performance across all conditions of our simulation study. In scenarios with sparse data coverage, performance increased with higher data availability in the given year and in adjacent years.

To illustrate the differences in the four latent trend models applied to real data, we present a case study with trajectories of political trust based on cross-national survey data from three selected countries: Hungary, Poland, and Spain, between 1995 and 2019. A comparison of political trust trajectories estimated by the four types of latent trend models points to further considerations for model choice, namely the amount of smoothing and the sensitivity of the estimated trajectory to single data points (in this case: single surveys), as well as the sensitivity of uncertainty estimates to data availability. Based on these results, GP, AR, and RW models may be better suited for estimating short-term volatility, on the condition that a sufficient amount of uniformly high-quality data is available. TPS may be more appropriate for estimating longer-term tendencies, especially if there is evidence of mixed data quality.

2. FRAMEWORK FOR MODELING AGGREGATE PUBLIC OPINION

The framework we propose consists of three components: the measurement model, the latent trend model, and sample representativeness corrections. The focus of this article is on the second element, the latent trend model, so we address the remaining two only briefly. The measurement model links the latent variable, the individual's unobservable level of a trait, to the observed survey responses. Common approaches in models estimating dynamic public opinion include individual- or group-level ordinal models (e.g., Caughey and Warshaw 2015; Solt 2020; Kołczyńska et al. 2020), binary models applied to dichotomized data (e.g., McGann 2014; Claassen 2019), and linear models applied to ordinal data treated as continuous (e.g., Durand et al. 2022). In the present analysis, we choose the variant of modeling individual-level data with ordinal item response theory (IRT) models, which has several advantages in comparison to the other listed options: it reduces information loss, respects the ordinal nature of the data, enables accounting for the different scale lengths, and enables the inclusion of individual-level predictors. Future studies may examine the relative advantages and disadvantages of these measurement models more systematically.

The latent trend model takes into account the temporal sequence of the data and—based on assumptions about the stickiness and other properties of societal public opinion—estimates a trend from the aggregated individual latent scores. We elaborate on the examined approaches to latent trend modeling in detail below. The most common strategy for correcting sample representativeness is weighting the data; alternatives include multilevel regression and poststratification (Gelman and Little 1997). Since our simulations do not involve non-response or individual-level covariates, we leave the relative merits of weights and poststratification to future research. While the framework was developed for studying public opinion, latent trend models are also applicable to analyses of other types of trajectories in areas as diverse as democracy (Treier and Jackman 2008) and lake sediments (Simpson 2018).

2.1 Measurement Model

The measurement model is formulated as an ordinal cumulative model (e.g., Samejima 1997; Bürkner and Vuorre 2019) of individuals' (ordinal) survey responses on a given item, that is, on the most fine-grained level of the survey data. The cumulative model assumes that the ordinal response y originates from the categorization of a latent, normally distributed, continuous variable $\tilde{y}$ according to an ordered latent threshold vector τ:

$$y = k \Longleftrightarrow \tau_{k-1} < \tilde{y} \leq \tau_k, \quad (1)$$

where $k = 1, \ldots, K + 1$ denotes the ordinal response categories implied by K internal thresholds $\tau = (\tau_1, \ldots, \tau_K)$ and two external thresholds $\tau_0 = -\infty$ and $\tau_{K+1} = \infty$. Assuming the latent factorization $\tilde{y} = \mu + \varepsilon$ into a latent mean μ to be predicted and a standard normally distributed error term ε implies that the probability of y being equal to k is given by

$$p(y = k|\mu) = \Phi(\tau_k - \mu) - \Phi(\tau_{k-1} - \mu),$$

where Φ is the cumulative distribution function of the standard normal distribution. This distributional assumption has the advantage that the latent scale is interpretable as (standardized) z-values, a widely known and understood scale, but other choices are possible as well (Bürkner and Vuorre 2019). The design of survey questions, in particular the length of response scales as well as question wording, varies across projects, and there are also possible differences between surveys within the same project. We use the term project to refer to the organization that coordinates the survey process and publishes the survey data under the same brand, for example, the European Social Survey or the World Values Survey. Within projects, surveys include data collected from the same sample of respondents in the same fieldwork process. Multi-wave projects include many waves of data collection. Within each project, the survey process is coordinated and standardized, a condition to achieve comparability of surveys from different countries and years.

Including project bias in models is intended to account for the differences in item design and other aspects of the survey process, which may affect respondents' answers. Because of the differences in item design, we assume a separate threshold vector for each survey-item combination. Further, to ensure that the latent trend model (see below) is jointly identified with the threshold vectors, we apply a sum-to-zero constraint to each of these threshold vectors. This constraint implies that changes in the latent trend over time are assumed to reflect actual attitude changes instead of changes to the content or interpretation of items.

The latent mean μ_i for observation i now is modeled by

$$\mu_i = b_0 + b_{\text{project}_i} + b_{\text{item}_i} + b_{\text{person}_i} + f(t_i), \quad (2)$$

where b_0 is the mean level of the attitude of interest across time, b_{project} and b_{item} are project- and item-specific deviations from the mean (with a sum-to-zero constraint across projects and items for identification), b_{person} is the person-specific deviation from the mean (with a soft sum-to-zero constraint via a hierarchical normal prior, Bürkner 2021), $f(.)$ is an unknown mean-zero function of time, and t_i represents the time point to which observation i belongs. This model can be understood as an additive multilevel model on the latent scale (Bürkner 2018). Importantly, the model of μ_i can easily be extended to more additive terms that might be relevant in a given survey context, thus providing researchers with a lot of flexibility to adjust the presented framework to their specific modeling challenges. The latent trend we are primarily interested in is given by $b_0 + f(t)$, while the other terms are added to account for the data structure commonly encountered in survey research.

2.2 Latent Trend Models

As latent trend models *f*, we consider four approaches to modeling non-linear associations that do not require the user to a priori specify their functional form: (a) thin-plate smoothing splines (Wood 2003), (b) zero-mean GPs with a smooth covariance kernel (Williams and Rasmussen 1996; Rasmussen and Williams 2006), (c) AR processes of order 1 (Box et al. 2015), and (d) RW processes (Box et al. 2015). Approaches (a) and (d) have been applied already in earlier research of modeling latent trends in public opinion across time with survey data (RW: Caughey and Warshaw 2015; Claassen 2019; Solt 2020; TPS: Kołczyńska et al. 2020) but have never been compared with each other. Approaches (b) and (c) have, to our knowledge, not yet been applied in this context.

2.2.1 Thin-plate splines

For a basic intuition, splines can be thought of as piecewise polynomials that are smoothly connected at their intersections, used for smoothing and interpolation (for an overview see, e.g., Wood 2017). As our first approach to modeling *f*, we consider a more advanced form of splines, penalized smoothing splines, which can be expressed via a basis function approach:

$$f(t) = \sum_{m=1}^{M} \alpha_m \phi_m(t) + \sum_{l=1}^{L} \beta_l \psi_l(t),$$

where ϕ_m and ψ_l are simple analytic basis functions (e.g., polynomials), and α_m and β_l are the corresponding spline coefficients (model parameters) to be estimated from the data. That is, for any given set of basis functions, the shape of *f* is defined by the spline coefficients. The conceptual difference between the coefficients sets $\{\alpha_m\}$ and $\{\beta_l\}$ is that one of them is penalized while the other is not (Wood 2004). Suppose α_m are the penalized coefficients, then, from a Bayesian perspective, this means that they have a joint hierarchical prior

$$\alpha_m \sim \text{prior}(\lambda)$$

with hyperparameters λ that are themselves model parameters estimated from the data; whereas β_l have mutually independent, potentially flat priors (Wood 2004). The here considered splines are parameterized in a way that the priors of α_m are zero-mean normal distributions with hyperparameters λ representing standard deviations over the non-linear space of the splines (Wood 2004). Importantly, λ are estimated from the data, along with all other model parameters. There is a wide range of basis functions to choose from implying different properties for the corresponding splines. In this article, we focus on (low-rank) TPS (Wood 2003). They constitute a class of smoothers that can be considered optimal solutions to the variational problem of balancing accuracy (fit to the observed data) and smoothness of the approximating function. More formally, they solve a class of minimization problems given by

$$||z - f(x)||^2 + \lambda J_{md}(f),$$

where z denotes the outcome variable to be predicted, $f(x)$ the corresponding predictions based on input variables x, and J_{md} is a penalty term depending on the number of input variables d and the desired degree of smoothness m (i.e., the number of minimally existing derivatives of f) controlling the overall wiggliness of the resulting function. For example, if the penalty term is given by $J_{md}(f) = \int f^{(m)}(x)^2 dx$, the optimal f would be a cubic spline (James et al. 2013). The general solution to the above variational problem is given in equation (7) and surrounding equations in Wood (2003). In our application, we had only one input variable (time) such that $d = 1$ and we set $m = 2$. Further, since the exact solutions may be quite high-dimensional in the number of basis functions and thus computationally demanding, a low-rank approximation was obtained following Wood (2003).

2.2.2 Gaussian processes

As our second approach to modeling f, we consider (zero-mean) GPs (see, e.g., Roberts et al. 2013; Görtler et al. 2019, for a conceptual introduction). From this perspective, f is a realization of an infinite dimensional normal distribution:

$$f \sim \text{normal}(0, C(\lambda))$$

where C is a covariance kernel with hyperparameters λ that defines the covariance between two function values $f(t_1)$ and $f(t_2)$ for two time points t_1 and t_2 (Rasmussen and Williams 2006). Similar to the different choices of the basis function for splines, different choices of the covariance kernel lead to different GPs. In this article, we consider the squared-exponential kernel defined as

$$C(\lambda) := C(t_1, t_2, \sigma, \gamma) := \sigma^2 \exp\left(-\frac{(t_1 - t_2)^2}{2\gamma^2}\right)$$

with hyperparameters $\lambda = (\sigma, \gamma)$, expressing the overall scale of GP and the length-scale, respectively (Rasmussen and Williams 2006). The advantages of this kernel are that it is computationally efficient and (infinitely) smooth making it a reasonable choice for the purposes of the present article. Here again, λ are estimated from the data, along with all other model parameters. For the finite set of $t_{\max}$ observed realization of f, that is, the finite time points at which $f(t)$ is evaluated in the training data, this implies a multivariate normal prior as

$$(f(t_1), f(t_2), \ldots, f(t_{\max})) \sim \mathrm{normal}(0, \mathrm{Cov}),$$

where the covariance matrix Cov has elements $\mathrm{Cov}_{ij} = C(t_i, t_j, \sigma, \gamma)$ according to the chosen kernel.

2.2.3 Autoregressive processes of order 1

As our third approach to modeling f, we consider an order 1 AR process model (Box et al. 2015) such that

$$f(t) = \theta_t \sim \mathrm{normal}(\rho\theta_{t-1}, \sigma)$$

for $t = 2, 3, \ldots$ and $f(1) = \theta_1 \sim \mathrm{normal}(0, \sigma)$, where θ_t are trend values of the AR process, ρ is the order 1 AR hyperparameter controlling the strength of the AR dependency, and σ is a standard deviation hyperparameter controlling the size of the discrete "jumps" from time t to time $t+1$. As for the other latent trend models, the hyperparameters are estimated from the data along with all other model parameters. To our knowledge, general AR processes have not yet been applied in the context of modeling trajectories in public opinion based on survey data. However, the special case of $\rho = 1$ known as RW process has been applied multiple times already (Caughey and Warshaw 2015; Claassen 2019; Solt 2020). In this study, we investigate both AR and RW processes as candidates for the latent trend model.

3. SIMULATION DESIGN

The simulations envision data from a single country from five projects (data sources) over a period of 25 years. In all projects, each survey includes three questions on the same three issues, but the length of response scales varies by project. Data simulation scenarios represent combinations of (a) the shape of the true latent trend, (b) the presence of between-person variation, and (c) data availability.

3.1 Shape of the True Latent Trend

We examine seven shapes of the latent trend. The first six represent combinations of three variants of the frequency of the trend, and two variants of the long-term tendency, as presented in figure 1. The three frequencies we examine

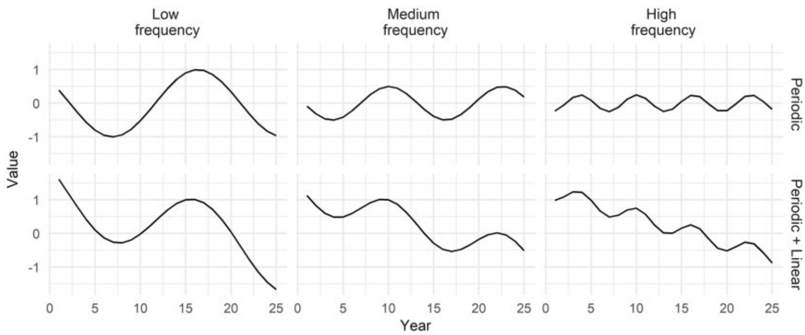

Figure 1. Shapes of the True Latent Trend.

include a pattern with low frequency and large amplitude (which covers much of the range of the latent scale), medium frequency and medium amplitude, and high frequency and small amplitude. All three trajectories represent realistic variation patterns of aggregate societal values or attitudes, of which each particular case of a country's trajectory likely constitutes a mix. Given that there is still little long-term longitudinal research of mass public opinion, the true volatility of the latent trend may not be known a priori. The two long-term tendencies include long-term stability and a weakly declining long-term linear trend, which we label "periodic" and "periodic + linear", respectively. Since we found no substantive differences between those two long-term scenarios, we focus only on the former in the presentation of results, while results for the "periodic + linear" scenarios are provided in the appendix in the supplementary data online. The seventh type of the latent trend is a "zero trend" where the true latent value equals zero in all years. While extended periods of no change in public opinion are unlikely to occur in reality, we use this scenario to verify, whether the models we use would overfit to noise and "invent" a pattern where there is none. The results for the "zero trend" scenarios are also presented in the appendix in the supplementary data online.

3.2 Between-Person Variation

When modeling latent trends in public opinion, it is often desirable to use responses to more than one question from the same respondent in a given survey. This nesting of responses in people needs to be accounted for by incorporating person-specific parameters (also known as person random effects) in the latent model, as shown in (2). We note that the standard practice in studies of trends in public opinion with survey data is to rely on group IRT models regardless of whether the same respondents contribute one or more responses, thus simply ignoring this aspect of data generation. In our simulations, between-person variation is generated from a normal distribution with mean 0 and standard deviation s. The two possibilities we consider include no between-person variation (i.e., $s = 0$) and $s = 1.2$, which is a realistic value based on the results from our empirical illustration with political trust. Comparing results for both sets of models demonstrates the effect of including between-person variation on model performance. In both conditions, the estimated models were correct, that is, they modeled between-person variation if included in the data generation process and did not otherwise.

3.3 Sparsity of the Data

In the full data scenario, data are available for all projects, items, and all 25 years. In the sparse data scenario, data are available for 31 out of the 125 project-years (25 percent), for 19 out of 25 years (76 percent), and without gaps in coverage longer than 1 year, as presented in figure 2. Whenever data

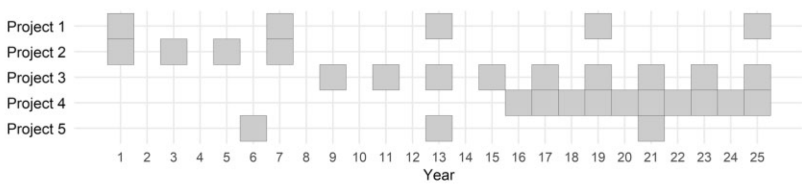

Figure 2. Data Availability Per Project and Year in the Sparse Scenario.

Table 1. Summary of Data Simulation Scenarios and Latent Trend Model Types

Data simulation scenarios			
Shape of the true trend	Between-person variation	Data sparsity	Latent trend models
High frequency, periodic	Zero	Full data	Thin-plate spline
Medium frequency, periodic	Non-zero	Sparse data	Gaussian process
Low frequency, periodic			Random walk model
High frequency, periodic + linear			Autoregressive model
Medium frequency, periodic + linear			
Low frequency, periodic + linear			
Zero (flat)			

NOTE.—On data simulated within each data simulation scenario, corresponding to combinations of the conditions in first three columns, we estimated four types of latent trend models, listed in the fourth column.

are available, they are available for all three items. In the light of the availability of survey data worldwide, this is a realistic but relatively comfortable situation characteristic for well-surveyed regions, including most of Europe.

Overall, we have $7 \times 2 \times 2 = 28$ data scenarios, which correspond to all combinations of the conditions in the three first columns in table 1. For each data scenario, we simulated 100 datasets.

3.4 Data Simulation

The simulated data include five survey projects, which—in the full data scenarios—provide data every year for 25 years. In each year's survey, there are 100 respondents who gave responses to three items that can be treated as indicators of an underlying latent variable. The number of respondents is lower than the typical sample size in cross-national surveys, which usually ranges between 1,000 and 2,000. The smaller number of respondents in our simulation study was chosen to control computation time. In our data, there is no unit or item non-response. Altogether, each simulated dataset in the full data scenarios has $5 \times 25 \times 100 \times 3 = 37,500$ observations. In the sparse data scenarios, each dataset has $31 \times 100 \times 3 = 9,300$ observations.

Data are simulated with the following procedure. For each of the true latent trends described in section 3.1:

(1) Draw project bias (five projects) and item bias (three items) from $\text{normal}(0, 1)$, and center the project and item biases to each have a mean of 0.
(2) For each survey, draw 100 person-specific deviations from $\text{normal}(0, s)$, where s equals 0 in the scenario without between-person variation, and s equals 1.2 in the scenario with between-person variation.
(3) Each respondent's latent mean μ_i per item is the sum of project and item bias, person-specific deviation, and latent trend $f(t)$, following (2).
(4) Draw per-observation random errors ε_i from $\text{normal}(0, 1)$.
(5) Each respondent's latent response $\tilde{y}_i$ per item is the sum of the latent mean and random error.
(6) For each project, draw the number of thresholds from the set of options $\{1, 3, 4, 6, 9, 10\}$ corresponding to response scale lengths represented in surveys $\{2, 4, 5, 7, 10, 11\}$ with probabilities of $\{0.1, 0.2, 0.2, 0.2, 0.2, 0.1\}$, which roughly reflects the prevalence of the response scale lengths in cross-national surveys.
(7) Draw the threshold values τ_k from $\text{normal}(0, 1)$.
(8) Center the thresholds to have a mean of 0 thereby enforcing the sum-to-zero constraint mentioned in section 2.1.
(9) Use the thresholds to cut the latent response into discrete categories following (1).

(10) In the sparse scenario, keep only the project-years as indicated in figure 2 and discard the remaining simulated data.

On each of the 2,800 datasets (100 datasets for each of 28 scenarios) we ran four types of latent trend models discussed in section 2.2.

3.5 Model Estimation

The analysis was performed using fully Bayesian estimation with Hamiltonian Monte Carlo as implemented in Stan (Carpenter et al. 2017). We set normal$(0, 3)$ priors on the thresholds and for the remaining parameters we used default priors as specified in the brms package (Bürkner 2017). We list the priors in the appendix in the supplementary data online.

We used default settings for adapt delta of 0.8 and maximum tree depth of 10, and set initial values to 0 on the transformed, unbounded scale. Each model was run with one chain comprising 3,000 iterations, of which 1,000 were for warm-up. Convergence was monitored based on the R-hat convergence diagnostic, bulk effective sample size, and tail effective sample size (Vehtari et al. 2021). The model code is available in the replication materials.

The large total number of models made the whole task computationally intensive. Within each scenario, the 100 models were ran in parallel, one core per model, on a high-performance computing cluster. Each set of 100 models took up to 960 core-hours for the full data scenarios and up to 240 core-hours for the sparse data scenarios.

For the subset of models with the "zero" trend, we monitored estimation times per model. Median values for models with TPS, GPs, AR, and RW models in each data condition were within an order of magnitude: between 1 and 2 hours for full data and between 19 and 35 minutes for sparse data conditions. It needs to be noted that similar estimation times, while observed in our analysis, should not be extrapolated to other settings, as the four types of latent trend models scale differently with the amount of data: TPS scale linearly with the number of observations, GPs scale cubicly with the number of time points, and AR and RW models scale linearly with the number of time points. Additionally, in our analysis, all models shared the same measurement model, which contributed to similar estimation times.

We performed the analysis in R (R Core Team 2018) using the brms package (Bürkner 2017), which provides a high-level interface to the probabilistic programming language Stan (Carpenter et al. 2017). TPS estimation was performed via the mgcv package (Wood 2003). For parallelization, we used the packages parallel, part of base R, foreach (Weston 2020), and doParallel (Microsoft Corporation and Weston 2019). We also used several tidyverse packages (Wickham et al. 2019) for data processing and visualization.

4. SIMULATION RESULTS

The purpose of the models is to estimate average levels of the latent attitude in each year, so our analysis focuses on these estimates only. Within each data scenario, the crucial comparisons are between the performance of TPS, GP, AR, and RW models, but we also discuss differences in overall performance between data scenarios. We report the results in terms of RMSE of latent trend estimates compared to the true values (square root of the average of squared differences between parameter values from each posterior draw and the true value), posterior standard deviation, and absolute bias (absolute value of the difference between the posterior mean and the true value), as well as coverage by 66 percent and 90 percent posterior credible intervals. It is worth noting that, in examining posterior credible intervals, we are interested in frequentist properties of Bayesian credible intervals, although we acknowledge that these credible intervals are not necessarily designed to include the true value in the nominally indicated percentage of cases.

We analyze the results both descriptively, by examining the distributions of the performance metrics across 100 simulations per scenario, as well as by fitting Bayesian regression models to obtain (a) estimates and uncertainties of average performance per scenario and (b) estimates and uncertainties of average performance depending on the number of available surveys within each sparse scenario. Details are provided in the appendix in the supplementary data online, which also includes results for additional performance metrics.

First of all, we notice that across all datasets in all scenarios and for all latent trend models, RMSE remains below 0.6. The average RMSE per scenario barely exceeds 0.2 for only one of them (sparse data, presence of between-person variation, and the low-frequency trend) and remains below 0.2 for the others. Given the scale of the latent trend (z-values), these values indicate good performance of all models.

Comparing model performance across scenarios, models estimated on data without between-person variation perform better compared to models estimated on data with between-person variation in terms of RMSE, posterior standard deviation, and absolute bias (compare facets in the first and second column of figure 3, and those in the third and fourth column). Furthermore, models estimated on full data perform better than models estimated on sparse data (compare facets in the first and third column of figure 3, and those in the second and fourth column). In sum, the difficulty of the scenario has direct implications for model performance. At the same time, differences between the shapes of the true trends are in most cases relatively small and depend on the latent trend model.

We now turn to comparisons of latent trend models within each of the 12 data scenarios presented in figures 3 and 4. AR and RW models perform quite similarly and tend to have the highest RMSE (with the exception of the scenario with full data, no between-person variation, and high-frequency trend;

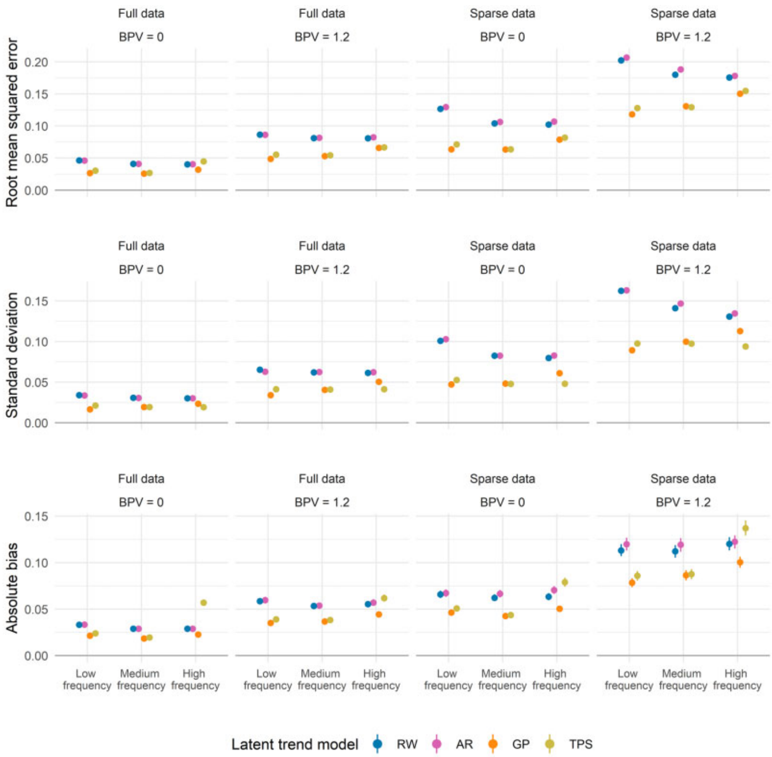

Figure 3. Estimated Average Levels of Root Mean Squared Error, Posterior Standard Deviation, and Absolute Bias by Scenario. Each point represents the estimated mean of the given performance metric across 100 models in each simulation scenario. Error bars indicate 95 percent posterior credible intervals of the mean estimates. Gray horizontal lines indicate no error. AR, autoregressive; BPV, between-person variation; GP, Gaussian process; high frequency, high-frequency trend; low frequency, low-frequency trend; medium frequency, medium-frequency trend; RW, random walk; TPS, thin-plate spline.

see facets in the top row of figure 3). The difference between AR and RW models and the other two is the biggest for the low-frequency trend, smaller in the medium-frequency scenario, and the smallest in the high-frequency scenario. RMSE for TPS and GP models is generally very similar.

The differences in RMSE between AR and RW models, and GP and TPS models are largely due to higher posterior standard deviations, and to a lesser extent due to absolute bias (figure 3). In terms of absolute bias, AR and RW models also perform worst with the exception of high-frequency trends. In the high-frequency scenarios, TPS models have the highest absolute bias.

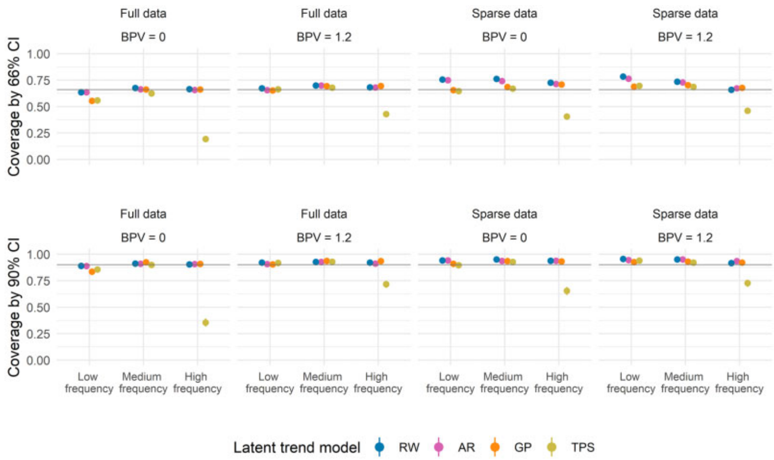

Figure 4. Estimated Average Levels of Coverage by Scenario. Each point represents the estimated mean level of coverage by the given posterior credible interval across 100 models in each simulation scenario. Error bars indicate 95 percent posterior credible intervals of the mean estimates. Gray horizontal lines indicate the reference value corresponding to the nominal width of the credible interval. AR, autoregressive; BPV, between-person variation; GP, Gaussian process; high frequency, high-frequency trend; low frequency, low-frequency trend; medium frequency, medium-frequency trend; RW, random walk; TPS, thin-plate spline.

In most cases, coverage is very close to what is expected based on the nominal width of the credible interval, with two main exceptions (cf. figure 4). First, TPS models exhibit lower coverage in models estimated on data with the high-frequency trend. The differences are substantial, with an average coverage of the 66 percent credible interval not exceeding 50 percent, and as low as under 20 percent in the scenario with full data and no between-person variation. Similar gaps are observed for other credible intervals. Second, coverage for AR and RW models tends to be higher than expected in the scenarios with sparse data. In scenarios with the low-frequency trend and sparse data, coverage by the 66 percent credible interval for AR and RW models reaches 80 percent. More generally, coverage by credible intervals for AR and RW models in scenarios with sparse data tends to exceed the credible intervals' nominal values.

We also examined the role of data availability on model performance in the sparse scenarios. Data availability is measured as the number of surveys available in the given and adjacent years and ranges from 1 in years 3, 9, and 11 to 7 in year 20 (cf. figure 2). In all scenarios having more surveys translates into lower RMSE, posterior standard deviation, and absolute bias (figure 5). More data also reduces over-coverage in AR and RW models, at least in the low-frequency scenarios. Perhaps more importantly, a higher number of available

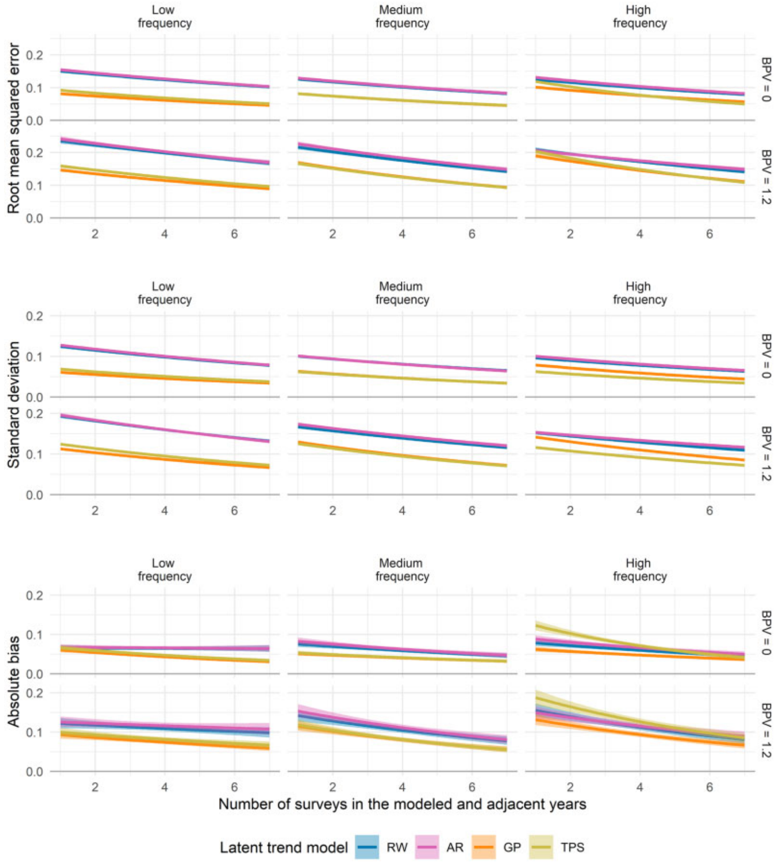

Figure 5. Estimated Values of Root Mean Squared Error, Posterior Standard Deviation, and Absolute Bias by the Number of Available Surveys in the Given and Adjacent Years for Each Sparse Data Simulation Scenario. Gray horizontal lines indicate the reference value of 0. Uncertainty is indicated as 95 percent posterior credible intervals. AR, autoregressive; BPV, between-person variation; GP, Gaussian process; high frequency, high-frequency trend; low frequency, low-frequency trend; medium frequency, medium-frequency trend; RW, random walk; TPS, thin-plate spline.

surveys substantially improves coverage in the high-frequency scenario in models with TPS, which on average suffer from under-coverage (figure 6).

From all these comparisons, GPs emerged as the option with the most reliable performance across metrics and data scenarios. Under conditions of our simulation study, from the performance point of view, GPs could be considered advisable if data coverage is incomplete and the shape of the latent trend is not known a priori. The results for the Periodic + Linear trends were similar to those of the Periodic trends. The zero trends exhibited substantially better performance, with RMSE, posterior standard deviation, and absolute bias

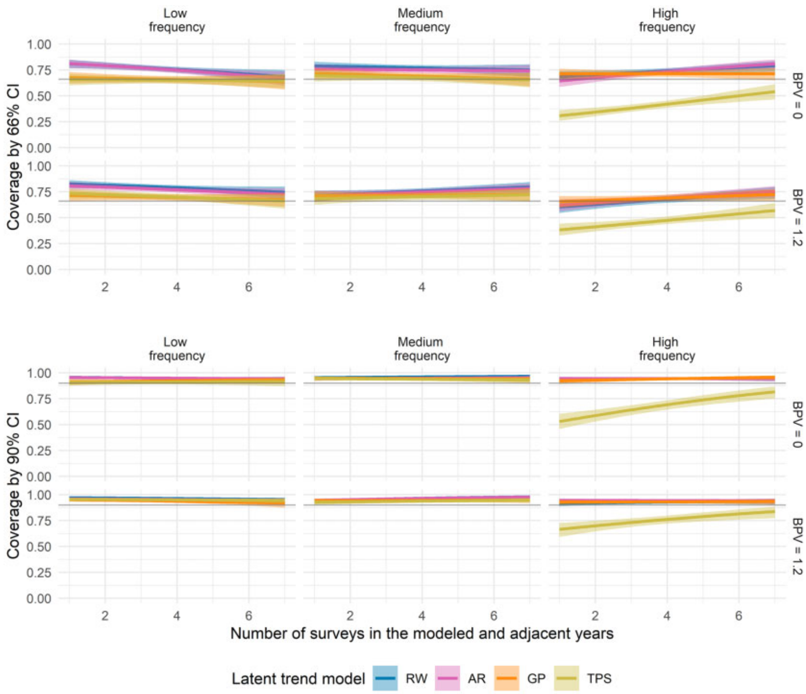

Figure 6. Estimated Values of Coverage by the 66 and 90 Percent Posterior Credible Interval by the Number of Available Surveys in the Given and Adjacent Years for Each Sparse Data Simulation Scenario. Gray horizontal lines indicate the reference value corresponding to the nominal width of the credible interval. Uncertainty is indicated as 95 percent posterior credible intervals. AR, autoregressive; BPV, between-person variation; GP, Gaussian process; high frequency, high-frequency trend; low frequency, low-frequency trend; medium frequency, medium-frequency trend; RW, random walk; TPS, thin-plate spline.

reaching about half of the respective levels in among models estimated on data in the Periodic scenarios. At the same time, in the zero trend scenarios, all models have a tendency for over-coverage, which is the largest in the case of AR models and the smallest in the case of TPS. In the sparse scenarios, this tendency is somewhat reduced in cases of higher data availability, especially for TPS models. Detailed results are presented in the appendix in the supplementary data online.

5. ILLUSTRATION: TRENDS IN POLITICAL TRUST

As a real-data application, we use the example of political trust measured with data from European cross-national survey projects: Candidate Countries

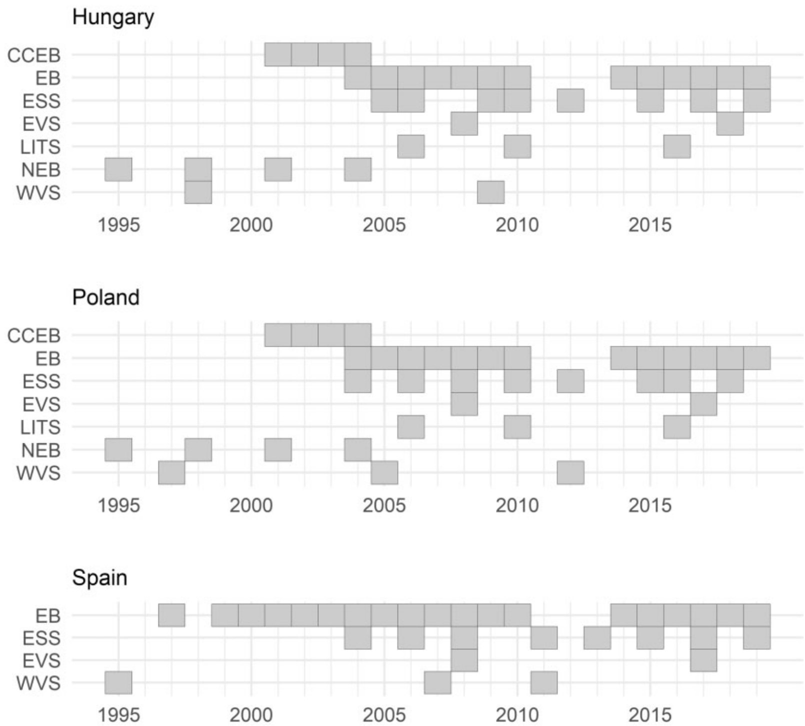

Figure 7. Data availability per project and year for Hungary, Poland, and Spain. CCEB, Candidate Countries Eurobarometer; EB, Eurobarometer; ESS, European Social Survey; EVS, European Values Study; LITS, Life in Transition Surveys; NEB, New Europe Barometer; WVS, World Values Survey.

Eurobarometer (CCEB), Eurobarometer (EB), European Social Survey (ESS), European Values Study (EVS), Life in Transition Survey (LITS), New Europe Barometer (NEB), and World Values Survey (WVS). From these projects, we selected surveys that include items on trust in the parliament, political parties, and the justice system, from three countries: Hungary, Poland, and Spain between 1995 and 2019. References to the data files, links to project websites with survey documentation, and the wording of survey items are provided in the appendix in the supplementary data online. Figure 7 illustrates the availability of survey data per project and country. In all three cases, there were fewer surveys in the 1990s compared to the later period, a pattern characteristic of most European countries, which will have consequences for model estimates.

For each country, we pooled the data from all projects, and on the resulting datasets, we estimated four models where time was modeled as a TPS, GP, and AR and RW processes. The models accounted for between-person variation,

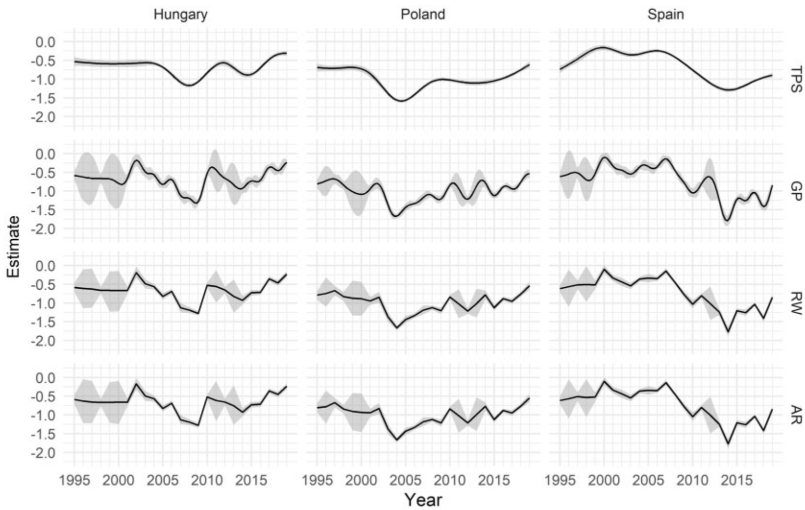

Figure 8. Estimated Trajectories of Political Trust in Hungary, Poland, and Spain, with Three Types of Time Trend Models: Thin-Plate Splines (TPS), Gaussian Processes (GP), Random Walk (RW), and Autoregressive (AR) Models. Shaded regions indicate 95 percent posterior credible intervals.

as well as project and item bias, similar to the models used in the simulation analysis. An evaluation of model fit, using leave-one-out cross-validation (LOO) as developed by Vehtari et al. (2016) and implemented in the loo package (Vehtari et al. 2020), indicates that RW models exhibit better fit than the remaining types, while TPS models consistently perform the worst. Details are provided in the appendix in the supplementary data online.

Estimated average levels of political trust as well as 95 percent posterior credible intervals of these estimates are presented in figure 8. These estimates show substantial agreement in the overall trajectories across models. At the same time, they reveal important differences in the degree of smoothing, as well as in the variation in the amount of uncertainty depending on data availability.

Trajectories estimated with TPS are the least volatile and capture only the long-term tendencies in political trust. The estimated uncertainty is rather small, with the average width of 95 percent posterior credible intervals of around 0.03 (on the probit scale) in all three countries. In periods with no data, the intervals are about three times as wide as in periods with data. Based on the simulation results, it is likely that the uncertainty is underestimated, especially when there is little data available.

Estimates from GP models are much more volatile and include short-term changes that are absent in the trajectories estimated with TPS. For periods without data, the estimated uncertainty is about 10 times as high as in periods

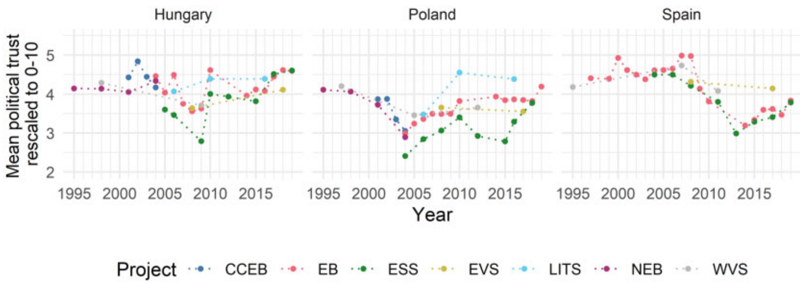

Figure 9. Averages of Trust in Parliament, Political Parties, and Justice System, Rescaled to 0–10 Range in Hungary, Poland, and Spain. Average political trust was calculated by rescaling the three trust variables to the 0–10 range, calculating mean levels of each trust type within country-year-projects, and averaging across the three trust means. CCEB, Candidate Countries Eurobarometer; EB, Eurobarometer; ESS, European Social Survey; EVS, European Values Study; LITS, Life in Transition Surveys; NEB, New Europe Barometer; WVS, World Values Survey.

with the most data. For example, in 1996–1997 and 1998–1999 in Hungary, two 2-year periods without any data, the width of posterior credible intervals was estimated at around 1.4 on the probit scale, which exceeds the range of the mean estimates of trust in this country.

The distinctive feature of AR and RW models is their discrete character. Estimates of these models are similar to those of the GPs with regard to the presence of short-term volatility and the sensitivity of uncertainty estimates to data availability. In periods with high data availability, the width of the credible intervals in AR and RW models is slightly smaller but overall similar to the GP models. In years with no data, credible intervals in AR and RW models are about two-thirds of those in the GP models. AR and RW models also estimate somewhat less volatility in political trust, based on the range of posterior means, compared to GP models.

Comparing model estimates with raw means of political trust rescaled to a common 0–10 range in figure 9 shows how GP and RW models reflect the local ups and downs in the survey data, while TPS smooth over much of that local variation. For example, the peak in Hungary in the year 2002, visible in the GP, AR, and RW models, reflects a single CCEB survey, in which average political trust was higher than in CCEB in adjacent years.

When comparing estimated trajectories from the four latent trend models, it is worth keeping in mind that their correspondence to the unknown true trajectory depends on the extent to which the model captures important features of the data-generating process. In the present analysis, the model accounts for project bias that is assumed to be constant across all project waves and does not allow for possible additional wave- or survey-specific bias. The sensitivity of the GP, AR, and RW models to single surveys is desirable when the goal is to estimate fine short-term fluctuations with data characterized by uniformly high quality of sample representation and measurement across all data sources. If this was true in the case of our survey data, we would, for example, expect agreement between projects in the trends of political trust. The projects presented in figure 9 exhibit broadly similar trends in political trust in the three countries, but also some differences. For example, in Spain, political trust according to EB increased between 2006 and 2008 but declined according to ESS. In Poland, EB data suggest the stability of political trust between 2010 and 2015, whereas ESS points to a decrease. Given these disagreements between data sources, the reliance of GP, AR, and RW models on single surveys to locally determine the trend may constitute over-fitting to data idiosyncrasies. This can also be viewed as a limitation of our measurement model, which did not allow for additional wave- or survey-specific bias, thus pushing some of the related signals onto the latent trend.

6. CONCLUSION

We introduced a framework to jointly model individual-level responses across items, surveys, projects, and time by combining ordinal measurement models with additive multilevel models on the latent scale. Within this framework, we compared four approaches to modeling latent trends that can be used in analyses of public opinion: TPS, GPs, AR, and RW models. Our simulation scenarios vary the shape of the true latent trend, the complexity of the model, and data availability. Overall, our results show that all four types of latent trend models perform well even in the most difficult scenarios with between-person variation, sparse data, and a trend with frequent small fluctuations.

Across our simulation scenarios, GP models (with squared exponential kernel) exhibited the most reliably high performance. TPS performed equally well, except for cases when the true trend was weak with frequent small fluctuations. In that latter case, TPS underestimates the uncertainty unless sufficient data were available; with enough data the under-coverage by TPS models was small. The comparably small uncertainty regions of the TPS may be the result of the underlying function not being wiggly enough, presumably due to (implicitly) fixing the length-scale parameter, which is common among spline approaches (Wood 2003; Kammann and Wand 2003). Viewed from a different perspective, the TPS represents a deterministic function of time, which may lead to auto-correlated residuals that cannot be fully captured by the model, again leading to a potential underestimation of uncertainty, depending on the nature of the true latent trend. Relatedly, extrapolation of a deterministic, data-driven trend model beyond the time range of the data is problematic, which is a drawback of this approach. AR and RW models tended to overestimate uncertainty, presumably because the AR and RW assume a discrete trend while, in our simulations, the true trend was a smooth function. Furthermore, AR and RW models on average had the highest RMSE of all latent model types. In the future, it could be interesting to study extensions of the RW prior to the residuals, such as AR-integrated-moving average priors (Kitagawa and Gersch 1984; Harvey 1989; Box et al. 2015), or special cases thereof, to increase the flexibility of the discrete time-series process, and perhaps improve accuracy and uncertainty calibration. Furthermore, it might be interesting to study combinations of splines and AR priors on the residuals to capture both the smooth latent trend and potentially remaining temporal dependencies.

The real-data illustration consisting in the estimation of political trust trajectories in three European countries based on data from cross-national surveys indicated superior predictive performance of RW models and inferior performance of TPS models. The analysis also reveals other differences between the four latent trend models: compared to TPS, the remaining three model types are much more sensitive to short-term fluctuations implied by the survey data, even if stemming from single surveys. The uncertainty estimates in GP, AR, and RW models were also substantially more sensitive to data gaps. The wide posterior credible intervals in years without any survey data, similar in width to the overall variation of the estimated trend in the given time series, in practice convey minimal information about the estimated quantity in these years. These results suggest that TPS may be more suitable for estimating trends of characteristics that—at least on the societal level—are known to change slowly, especially if the available survey data are noisy and have incomplete coverage. GP, AR, and RW models may be better suited when modeling more volatile attitudes but require higher quality data and time series without data gaps. That said, as already mentioned earlier, improving the measurement model to capture additional sources of variation could prevent the related signal to be absorbed into the latent GP, AR, and RW trend and thus reduce their sensitivity to local data idiosyncrasies. In other words, one cannot separate the latent model from the measurement model due to their complex interactions, and potential misspecifications in one of the two models are likely to affect both of them.

In addition to considering these differences, the ultimate choice of the latent trend model would likely depend on other factors as well, including the model's ability to incorporate individual-level predictors, avoiding possible problems with model specification or estimation, and on the number of observations and time points due to the latent models' different scaling. Regardless of which model is chosen, we emphasize the need to account for the uncertainty associated with estimates of the latent trends in explanatory models.

The framework for analyzing public opinion trends with data from different survey projects conceptually connects to individual data meta-analysis performed in particular in medical or psychological research, where individual participants' data from different studies are pooled and analyzed jointly (e.g., Riley et al. 2010; Riley 2010; Ioannidis 2017). In social survey research, methods for combining existing survey datasets are studied under the label "ex-post survey data harmonization" (Wolf et al. 2016). Despite the differences in the types of data and analyses, both lines of research emphasize the need to take into account the clustering of participants within studies as well as to include study- and participant-level characteristics.

6.1 Limitations

The simulations we designed take into account some of the issues that researchers of aggregate public opinion encounter, most notably related to data availability. Others have remained unexamined, which limits the generalizability of our results. Such unexamined conditions include possible problems with data quality, such as representation errors, added complexity of the measurement model, such as varying discrimination parameters across items and/or projects, and other true latent trends than the ones we considered, in particular trends that result from random processes such as GPs or discrete AR models. All these aspects will influence the performance of the estimated latent trend models, and their suitability for a particular application.

We also examined just one sparsity pattern, due to space limitations, and one that is rather optimistic in the light of survey data availability on the issues that are of interest to social scientists (cf. Claassen 2020). When including item and project bias, what will matter is not only the number of surveys overall, but the number of surveys per project and how the surveys from different projects are spread over the period of interest. In applications with real data, the combination of the sparsity pattern and the true latent trend will likely determine the difficulty of the modeling task, because detecting smaller fluctuations requires more and better data. Overall, when dealing with sparse data coverage—which is the reality in modeling public opinion over time—caution is advised when drawing conclusions about fluctuations smaller and more frequent than the ones in our high-frequency scenario.

SUPPLEMENTARY MATERIALS

Supplementary materials are available online at academic.oup.com/jssam.

AUTHORS' CONTRIBUTIONS

Marta Kołczyńska and Paul-Christian Bürkner designed the simulation analysis, wrote the analysis code, performed the analysis, wrote the original, and revised texts. Paul-Christian Bürkner designed the analysis methods and oversaw the analyses. Marta Kołczyńska prepared the case study.

Data Availability

The replication code for this article is available in the Open Science Framework repository: https://osf.io/j45er/.

REFERENCES

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015), *Time Series Analysis: Forecasting and Control*, Hoboken: John Wiley & Sons.

Bürkner, P.-C. (2017), "Brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, 80, 1–28. doi:10.18637/jss.v080.i01

———. (2018), "Advanced Bayesian Multilevel Modeling with the R Package Brms," *The R Journal*, 10, 395–411.

———. (2021), "Bayesian Item Response Modelling in R with Brms and Stan," *Journal of Statistical Software*, 100, 1–54.

Bürkner, P.-C., and Vuorre, M. (2019), "Ordinal Regression Models in Psychology: A Tutorial," *Advances in Methods and Practices in Psychological Science*, 2, 77–101.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. doi:10.18637/jss.v076.i01

Caughey, D., and Warshaw, C. (2015), "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model," *Political Analysis*, 23, 197–211.

Claassen, C. (2019), "Estimating Smooth Country–Year Panels of Public Opinion," *Political Analysis*, 27, 1–20.

———. (2020), "Does Public Support Help Democracy Survive?" *American Journal of Political Science*, 64, 118–134.

Durand, C., Peña Ibarra, L. P., Rezgui, N., and Wutchiett, D. (2022), "How to Combine and Analyze All the Data from Diverse Sources: A Multilevel Analysis of Institutional Trust in the World," *Quality & Quantity*, 56, 1755–1797.

Gelman, A., and Little, T. C. (1997), "Postratification into Many Categories Using Hierarchical Logistic Regression," *Survey Methodology*, 23, 127–135.

Görtler, J., Kehlbeck, R., and Deussen, O. (2019), "A Visual Exploration of Gaussian Processes," *Distill*, 4. doi:10.23915/distill.00017

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

Ioannidis, J. (2017), "Next-Generation Systematic Reviews: Prospective Meta-Analysis, Individual-Level Data, Networks and Umbrella Reviews," *British Journal of Sports Medicine*, 51, 1456–1458.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning* (Vol. 112), New York, NY: Springer.

Kammann, E. E., and Wand, M. P. (2003), "Geoadditive Models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 1–18.

Kitagawa, G., and Gersch, W. (1984), "A Smoothness Priors–State Space Modeling of Time Series with Trend and Seasonality," *Journal of the American Statistical Association*, 79, 378–389.

Kołczyńska, M., Bürkner, P.-C., Kennedy, L., and Vehtari, A. (2020), *Modeling Public Opinion Over Time and Space: Trust in State Institutions in Europe, 1989-2019*, SocArXiv. doi:10.31235/osf.io/3v5g7.

McGann, A. J. (2014), "Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm," *Political Analysis*, 22, 115–129.

Microsoft Corporation, and Weston, S. (2019), "doparallel: Foreach Parallel Adaptor for the 'Parallel' Package [Computer Software Manual]." R package version 1.0.15. Available at https://CRAN.R-project.org/package=doParallel.

R Core Team (2018), *R: A Language and Environment for Statistical Computing [Computer Software Manual]*. Austria: Vienna. Available at https://www.r-project.org/.
Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press.
Riley, R. D. (2010), "Commentary: Like It and Lump It? Meta-Analysis Using Individual Participant Data," *International Journal of Epidemiology*, 39, 1359–1361.
Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010), "Meta-Analysis of Individual Participant Data: Rationale, Conduct, and Reporting," *BMJ*, 340, c221.
Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013), "Gaussian Processes for Time-Series Modelling," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 1–25.
Samejima, F. (1997), "Graded Response Model," in *Handbook of Modern Item Response Theory*, eds. W. J. van der Linden and R. K. Hambleton, New York, NY: Springer, pp. 85–100.
Simpson, G. L. (2018), "Modelling Palaeoecological Time Series Using Generalised Additive Models," *Frontiers in Ecology and Evolution*, 6, 149.
Solt, F. (2020), *Modeling Dynamic Comparative Public Opinion*, SocArXiv. doi:10.31235/osf.io/d5n9p.
Treier, S., and Jackman, S. (2008), "Democracy as a Latent Variable," *American Journal of Political Science*, 52, 201–217.
Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020), "loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models." R package version 2.4.1. Available at https://CRAN.R-project.org/package=loo.
Vehtari, A., Gelman, A., and Gabry, J. (2017), "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC," *Statistics and Computing*, 27, 1413–1432. doi:10.1007/s11222-016-9696-4.
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021), "Rank-Normalization, Folding, and Localization: An Improved Rhat for Assessing Convergence of MCMC," *Bayesian Analysis*, 16, 1–28.
Weston, S.; Microsoft (2020), "foreach: Provides Foreach Looping Construct [Computer Software Manual]". R package version 1.5.0. Available at https://CRAN.R-project.org/package=foreach.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019), "Welcome to the Tidyverse," *Journal of Open Source Software*, 4, 1686.
Williams, C. K. I., and Rasmussen, C. E. (1996), "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems* (Vol. 8), eds. D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Cambridge, MA: MIT Press, pp. 514–520.
Wolf, C., Schneider, S. L., Behr, D., and Joye, D. (2016), "Harmonizing Survey Questions between Cultures and over Time," in *The SAGE Handbook of Survey Methodology*, eds. C. Wol, D. Joye, T. W. Smith, and Y.-c. Fu. London: SAGE, pp. 502–524. doi:10.4135/9781473957893.
Wood, S. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 95–114.
———. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686.
———. (2017), *Generalized Additive Models: An Introduction with R*, Boca Raton: Chapman and Hall/CRC.