Journal of Computational Science 80 (2024) 102326

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Bayesian inference for a spatio-temporal model of road traffic collision data

Nicola Hewett [a,*], Andrew Golightly [b], Lee Fawcett [a], Neil Thorpe [c]

[a] School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
[b] Department of Mathematical Sciences, Durham University, Stockton Road, Durham, DH1 3LE, UK
[c] Jacobs, Rotterdam House, 116 Quayside, Newcastle Upon Tyne, NE1 3DY, UK

ARTICLE INFO

Keywords:
Dynamic linear model (DLM)
Bayesian inference
Forward filter backward sampler
Markov chain Monte Carlo

ABSTRACT

Improving road safety is hugely important with the number of deaths on the world's roads remaining unacceptably high; an estimated 1.35 million people die each year (WHO, 2020). Current practice for treating collision hotspots is almost always reactive: once a threshold level of collisions has been exceeded during some predetermined observation period, treatment is applied (e.g. road safety cameras). However, more recently, methodology has been developed to predict collision counts at potential hotspots in future time periods, with a view to a more proactive treatment of road safety hotspots. Dynamic linear models provide a flexible framework for predicting collisions and thus enabling such a proactive treatment. In this paper, we demonstrate how such models can be used to capture both seasonal variability and spatial dependence in time dependent collision rates at several locations. The model allows for within- and out-of-sample forecasting for locations which are fully observed and for locations where some data are missing. We illustrate our approach using collision rate data from 8 Traffic Administration Zones in the US, and find that the model provides a good description of the underlying process and reasonable forecast accuracy.

1. Introduction

Every year the lives of approximately 1.3 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury [1]. Working with collision counts can introduce issues of zero-inflation, especially over short time-frames. By working with rates over zones, we have the advantage of fewer zeros in the data set and upon removing these, we may treat the data as continuous, which can be mathematically convenient in terms developing a tractable model. Most road traffic data are recorded sequentially over time and it is common for there to be dependencies between each observation. Hence, it is necessary to account for these dependencies in the model via a time-series model, such as a state–space model. The use of state–space models in road safety analysis is relatively new and uncommon, though they provide advantages for prediction.

State–space models can be used for modelling univariate or multivariate time-series in the presence of non-stationarity, structural changes and irregular patterns (see e.g. [2,3]). Time-series analysis typically begins with the formulation of a model that accounts for temporal dependence, for example through auto-correlation, trend or seasonality. The use of state–space models within a time-series setting allows for uncertainty quantification in both the observation process and any dynamic variables that are not observed directly. Forecasting therefore accounts for these different sources of uncertainty and, when inferences are made within the Bayesian paradigm, additional parameter uncertainty. Throughout, we focus on a particular class of state–space model within which the observation and system equations involve linear functions of the latent process. Such models are known as dynamic linear models (DLMs, see e.g. [3,4]) and offer several practical benefits over their nonlinear counterparts. Notably, they admit a tractable observed data likelihood function, allowing a computationally efficient approach to inference and forecasting.

Gamerman and Migon [5] give a list of hierarchical dynamic linear models (DLMs) used for the state evolution, smoothing and filtering through the stages of the hierarchy. Previous applications of (hierarchical) state–space models in the road safety context have focused on temporal models of short-term travel time on a freeway stretch [6], road traffic collision analysis and prediction (see e.g [7–9]) and the impact of speed limit policies [10]. Despite their flexibility in describing the underlying data generating mechanism, state–space models and DLMs in particular, have been to date rarely exploited in the road safety context (see e.g. [6,11]).

Our contribution is a joint spatio-temporal model of collision rates over multiple zones. A DLM is used at the level of a single zone, and

* Corresponding author.
E-mail address: nicola.hewett@newcastle.ac.uk (N. Hewett).

https://doi.org/10.1016/j.jocs.2024.102326
Received 30 January 2024; Received in revised form 16 April 2024; Accepted 14 May 2024
Available online 17 May 2024

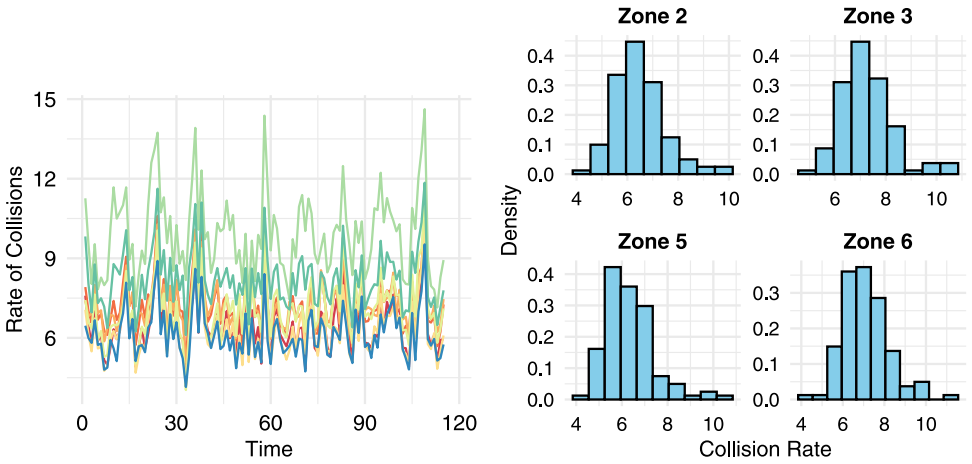

Fig. 1. Left: Time series plots of monthly collision rate in each of the 8 zones. Right: Histograms of collision rates for zones 2,3,5,6.

allows for seasonality via a single harmonic with time-varying amplitude and phase parameters. We then account for spatial dependence at nearby locations by adding a spatial Gaussian process to the system equation, thereby smoothing spatial deviations from the underlying temporal process. A similar approach in an environmental setting can be found in [12] (see also [13]). The resulting model allows for both within- and out-of-sample forecasting for locations which are fully observed and for locations at which some data are missing. A Bayesian approach is used to infer both dynamic and static model components and leverages the tractability of the observed data likelihood, which can be efficiently computed via a forward filter (see e.g. [14,15]). We apply the inference scheme to data consisting of monthly collision rates from neighbouring Traffic Administration Zones across a region of Florida/Georgia, USA. We assess the assumption of time-varying parameters governing the seasonal component to each zone separately before considering a joint model of all zones.

The remainder of this paper is organised as follows. A brief description of the data is given in Section 2. The structure of the DLM for a single zone and joint zones is given in Section 3. In Section 4 we outline the details of the Bayesian inference scheme, before considering the data application in Section 5. Conclusions are drawn in Section 6.

2. Data

We consider monthly collision rate data for neighbouring Traffic Administration Zones in North Florida/Southern Georgia, USA. For 8 Traffic Administration Zones we have monthly collision rates; for each zone we have 115 months of observations, the most recent being from April 2014. Fig. 1 shows the multiple data streams over time for the different zones. For all zones, the monthly collision rates exhibit sinusoidal patterns over a 12 month period. Histograms of the monthly collision rates suggest that a Gaussian observation model may adequately describe the observation process. Through scatter plots, we determined that there was clear temporal dependence between certain months in year t to year $t+1$, precluding the use of a simpler model with "month" as a fixed effect. Furthermore, zones geographically closer are more strongly correlated (see Fig. 2).

3. Dynamic linear model (DLM)

State–space models build on the relatively simple dependence structure of a (first order) Markov chain (in that information about some state θ_{t_i} carried by all previous values of the chain up to time t_{i-1} is the same as that carried by $\theta_{t_{i-1}}$ alone). They are made of two main components, observed data $(x_{t_1}, \dots, x_{t_n})$ and unobserved/latent states $(\theta_{t_0}, \dots, \theta_{t_n})$. Fig. 3 shows the evolution of a simple univariate state–space in which the continuous valued latent state process $\{\theta_{t_0}, \theta_{t_1}, \dots, \theta_{t_{i-1}}, \theta_{t_i}, \dots\}$ evolves according to a first order Markov chain with transition density $\pi(\theta_{t_i}|\theta_{t_{i-1}})$. The continuous-valued observation process $\{x_{t_1}, x_{t_2}, \dots, x_{t_{i-1}}, x_{t_i}, \dots\}$ is linked to the latent state process at an arbitrary time t_i via the density $\pi(x_{t_i}|\theta_{t_i})$; here it is assumed that the observed data are conditionally independent given the latent states. The observable process $\{X_{t_i}\}$ depends on the underlying, unobservable latent state process $\{\theta_{t_i}\}$ and we can reasonably assume that the observation X_{t_i} only depends on the state of the system at the time the measurement is taken, θ_{t_i}. It remains that we specify the relationship between X_{t_i} and θ_{t_i}, and between θ_{t_i} and $\theta_{t_{i-1}}$. In each case, we adopt linear relationships, and further assume that the errors in the state and observed components are independent and normally distributed. This structure leads to a dynamic linear model (DLM), given by the following equations:

$$\text{Observation Equation}: \quad X_{t_i} = F_{t_i}\theta_{t_i} + \nu_{t_i}$$
$$\text{System Equation}: \quad \theta_{t_i} = G_{t_i}\theta_{t_{i-1}} + \omega_{t_i}$$

Here, X_{t_i} is a scalar, θ_{t_i} is a $p\times 1$ vector, F_{t_i} is a $1\times p$ vector, G_{t_i} is a $p\times p$ matrix and $\nu_{t_i} \sim \mathrm{N}(0, V_{t_i})$ and $\omega_{t_i} \sim \mathrm{N}(0, W_{t_i})$ are independent white noise processes with known variance matrices V_t and W_t, typically assumed to be constant. Assuming that the initial latent state follows a Gaussian distribution gives

$$\begin{aligned}\theta_0 &\sim \mathrm{N}(m_0, C_0)\\ \theta_{t_i}|\theta_{t_{i-1}} &\sim \mathrm{N}(G_{t_i}\theta_{t_{i-1}}, W_{t_i})\\ X_{t_i}|\theta_{t_i} &\sim \mathrm{N}(F_{t_i}\theta_{t_i}, V_{t_i})\end{aligned}$$

for suitably chosen hyperparameters m_0 and C_0. In what follows we consider a DLM appropriate for data at a single zone, before considering a joint model over all zones.

3.1. Zone specific model

The data set described in Section 2 showed seasonality in that, over all zones there was a clear sinusoidal pattern about the rate of collisions over a year. Therefore, to account for this within the DLM we include a single harmonic. Note that it is possible to account for seasonality through the inclusion of multiple harmonics in the system equation (see e.g. [4]), however, we find that using a single harmonic and allowing the amplitude and phase to vary over time, provides a parsimonious modelling approach.

Consider first a single location. We assume constant variance matrices V and W and data at irregularly spaced times $t_1, t_2, \dots, t_n$. The observation equation is

$$X_{t_i} = F_{t_i}\theta_{t_i} + \nu_{t_i}, \qquad \nu_{t_i} \overset{indep}{\sim} \mathrm{N}(0, V), \tag{1}$$

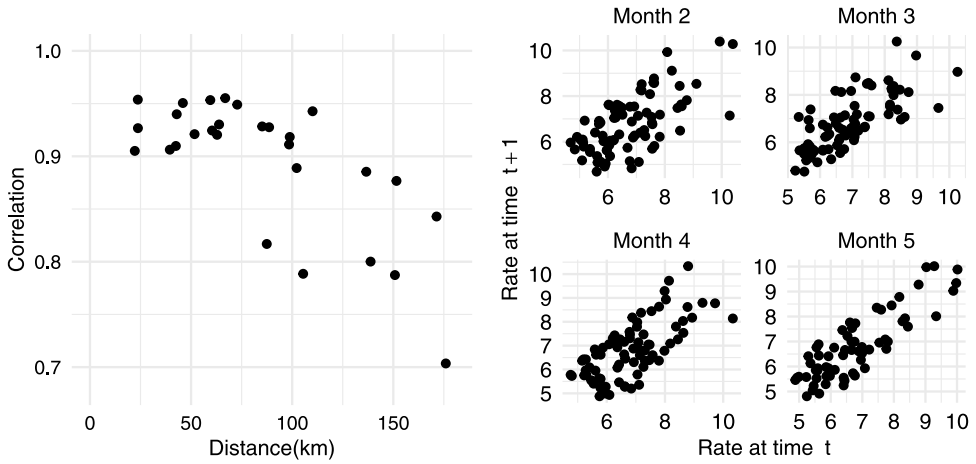

Fig. 2. Left: Correlation between the 8 zones against distance between zones (km). Right: The temporal dependence between observations in months 2,3,4,5 in consecutive years across all zones.

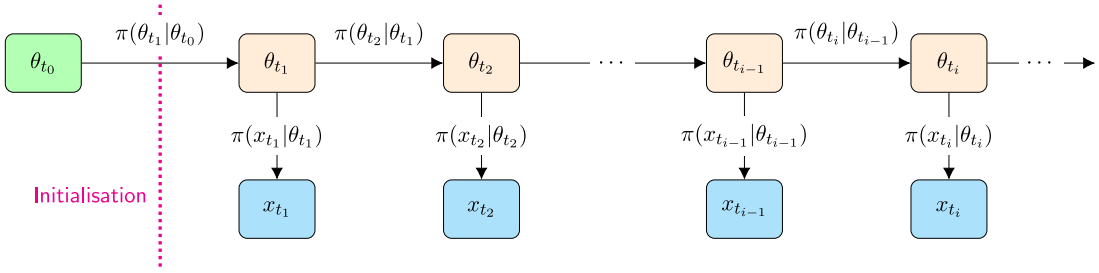

Fig. 3. Directed acyclic graph showing the dependence structure of the state–space model.

where $\theta_{t_i} = (\theta_{1,t_i}, \theta_{2,t_i}, \theta_{3,t_i})^T$ and the observation matrix is given by

$$F_{t_i} = \left(\sin\left(\frac{2\pi i}{P_x}\right), \cos\left(\frac{2\pi i}{P_x}\right), 1\right)$$

where P_x is the time corresponding to one complete period ($P_x = 12$ for seasonal data). Note that the observation equation can be written as

$$X_{t_i} = \tilde{\theta}_{1,t_i} \cos\left(\frac{2\pi i}{P_x} - \tilde{\theta}_{2,t_i}\right) + \theta_{3,t_i} + \nu_{t_i} \tag{2}$$

where the dynamic parameters in Eqs. (1) and (2) are related using

$$\tilde{\theta}_{1,t_i} = \sqrt{\theta_{1,t_i}^2 + \theta_{2,t_i}^2}, \qquad \tilde{\theta}_{2,t_i} = \tan^{-1}\left(\frac{\theta_{1,t_i}}{\theta_{2,t_i}}\right). \tag{3}$$

We impose some smoothness in these dynamic parameters by taking the system equation to be of the form

$$\theta_{t_i} = \theta_{t_{i-1}} + k_{t_i}\omega_{t_i}, \qquad \omega_{t_i} \overset{indep}{\sim} \mathrm{N}(0, W)$$

which has been further altered to allow for measurements that are irregularly spaced on a temporal grid. That is, we include a coefficient, k_{t_i}, in the variance in the state equation such that $k_{t_i}^2 = t_i - t_{i-1}$. Hence, the sinusoidal form DLM captures seasonality via a single harmonic whose amplitude $\tilde{\theta}_{1,t_i}$ and phase $\tilde{\theta}_{2,t_i}$ vary according to two transformed independent random walk processes. Other approaches for capturing seasonality include the use of multiple harmonics in the system equation (see e.g. [4]). A comparison of this approach versus the use of a single harmonic with time-varying amplitude and phase can be found in [12], who find that the latter offers a good compromise between model fit and computational cost.

3.2. Joint model over zones

We now consider a model of monthly collision rates that captures both the seasonality, and additionally, the correlation between nearby zones. Let $X_{t_i} = (X_{t_i}^1, \ldots, X_{t_i}^{n_z})^T$ denote the collection of monthly collision rates at time t_i with $X_{t_i}^j$ corresponding to zone j, and $j = 1, \ldots, n_z$. In Section 5.1 we find that amplitude and phase are plausibly constant for each zone. Therefore, for ease of notation, we fix $\theta_{1,t_i} = \theta_1$ and $\theta_{2,t_i} = \theta_2$ in what follows. The model at zone j is

$$X_{t_i}^j = \theta_1^j \sin\frac{\pi t_i}{6} + \theta_2^j \cos\frac{\pi t_i}{6} + \theta_{3,t_i}^j + \nu_{t_i}^j, \qquad \nu_{t_i}^j \overset{indep}{\sim} N(0, V^j),$$
$$\theta_{3,t_i}^j = \theta_{3,t_{i-1}}^j + k_{t_i}\omega_{t_i}^j + p_{t_i}^j, \qquad \omega_{t_i}^j \overset{indep}{\sim} N(0, W^j),$$

To induce correlation between nearby zones, we include the term $p_{t_i}^j$ as a component of a spatially smooth error process $p_{t_i} = (p_{t_i}^1, \ldots, p_{t_i}^{n_z})^T$. We model $\{p_{t_i}, t_i = 1, \ldots, n\}$ using independent (over i) zero-mean Gaussian processes so that

$$p_{t_i} \overset{indep}{\sim} GP\{0, f_3(\cdot; \eta_3)\}.$$

We impose smoothness by taking a squared exponential kernel for the covariance function. Hence, the covariance between spatial errors at locations j and j' is

$$f_3(d_{jj'}; \eta_3) = \mathrm{Cov}(\theta_{3,t_i}^j, \theta_{3,t_i}^{j'}) = \sigma_3^2 \exp(-\phi_3 d_{jj'}), \tag{4}$$

with $\eta_3 = (\sigma_3, \phi_3)^T$ parameterising the kernel; note that ϕ_3 determines the decay ratio of the correlation as the distance between sites j and j' ($d_{jj'}$) increases [16]. Similarly, we adopt GP priors for $\theta_1 = (\theta_1^1, \ldots, \theta_1^{n_z})^T$ and $\theta_2 = (\theta_2^1, \ldots, \theta_2^{n_z})^T$ so that $\theta_1 \sim GP(m_1(\cdot), f_1(\cdot; \eta_1))$ and $\theta_2 \sim GP(m_2(\cdot), f_2(\cdot; \eta_2))$ with f_1 and f_2 defined analogously to Eq. (4) with the addition of $m_1(\cdot)$ and $m_2(\cdot)$ as appropriate mean functions. Hence, the full spatial DLM model (over all locations) is

$$X_{t_i} = F_{t_i}\theta_{t_i} + \nu_{t_i}, \qquad \nu_{t_i} \overset{indep}{\sim} \mathrm{N}(0, \mathrm{diag}\{V^1, \ldots, V^{n_z}\}),$$
$$\theta_{3,t_i} = \theta_{3,t_{i-1}} + k_{t_i}\omega_{t_i}, \qquad \omega_{t_i} \overset{indep}{\sim} \mathrm{N}(0, \mathrm{diag}\{W^1, \ldots, W^{n_z}\} + K_3),$$

where $F_{t_i} = \text{diag}(F_{t_1}^1, \ldots, F_{t_i}^{n_z})$, $\theta_{3,t_i} = (\theta_{3,t_i}^1, \ldots, \theta_{3,t_i}^{n_z})^T$, $\theta_{t_i} = (\theta_1^1, \theta_2^1, \theta_{3,t_i}^1, \ldots, \theta_1^{n_z}, \theta_2^{n_z}, \theta_{3,t_i}^{n_z})^T$ and K_3 is an $n_z \times n_z$ matrix with (i,j)th element $f_3(d_{ij}, \eta_3)$.

4. Bayesian inference

For simplicity, suppose we have n_z zones with n observations in each zone. Let $V = (V^1, \ldots, V^{n_z})^T$ and $W = (W^1, W^2, \ldots, W^{n_z})^T$. Furthermore, let $\eta_3 = (\sigma_3, \phi_3)^T$ denote the hyperparameters governing $f_3(\cdot)$, with $\eta_1 = (\sigma_1, \phi_1)^T$ and $\eta_2 = (\sigma_2, \phi_2)^T$ denoting the hyperparameters governing $f_1(\cdot)$ and $f_2(\cdot)$ respectively. Let $x^j = (x_{t_1}^j, \ldots, x_{t_n}^j)^T$ denote the vector of collision rates at site j so that $x = (x^1, \ldots, x^{n_z})$ denotes the complete data set over all zones. The joint posterior over all dynamic and static parameters is proportional to the marginal static parameter posterior multiplied by the conditional posterior of the dynamic process $\theta_3 = (\theta_{3,t_0}, \ldots, \theta_{3,t_n})$ such that

$$\pi(\theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3, \theta_3 | x) \propto \pi(\theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3 | x) \times \pi(\theta_3 | \theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3, x).$$

Let ψ denote all fixed model parameters. To simulate realisations from the joint posterior we use a two step approach:

1. Simulate from the marginal posterior $\psi \sim \pi(\psi | x)$.
2. Simulate from the conditional posterior $\theta_3 \sim \pi(\theta_3 | \psi, x)$.

For step 1, as the marginal static parameter posterior is intractable, we use Markov chain Monte Carlo (see e.g. [17]). For step 2 we use a forward filter backward sampling (FFBS) algorithm (see e.g. [3]) to directly draw from $\pi(\theta_3 | \psi, x)$. We provide details as follows.

4.1. Simulation based inference

Let $\theta_{3,t_{0:n}} = (\theta_{3,t_0}, \theta_{3,t_1}, \ldots, \theta_{3,t_n})$ denote the collection of latent states up to time t_n and let $x = x_{t_{1:n}} = (x_{t_1}, \ldots, x_{t_n})$ denote the observed data. Note that $\theta_{3,t_i} = (\theta_{3,t_i}^1, \ldots, \theta_{3,t_i}^{n_z})^T$ and $x_{t_i} = (x_{t_i}^1, \ldots, x_{t_i}^{n_z})^T$. Upon assuming an independent prior specification for the constituent terms of ψ, Bayesian inference may proceed as follows. Integrating out the dynamic parameters, gives us the marginal posterior:

$$\pi(\theta_1, \theta_2, V, W, \eta_1, \eta_2, \eta_3 | x) \propto \pi(\theta_1 | \eta_1)\pi(\theta_2 | \eta_2) \left[\prod_{j=1}^{n_z} \pi(V^j)\pi(W^j) \right] \times \pi(\eta_1)\pi(\eta_2)\pi(\eta_3) \times \pi(x | \theta_1, \theta_2, \eta_3, V, W)$$

where the marginal likelihood $\pi(x | \theta_1, \theta_2, \eta_3, V, W)$ is given by

$$\pi(x | \theta_1, \theta_2, \eta_3, V, W) = \pi(x_{t_1} | \theta_1, \theta_2, \eta_3, V, W) \prod_{i=1}^{n-1} \pi(x_{t_{i+1}} | x_{t_{1:i}}, \theta_1, \theta_2, \eta_3, V, W) \tag{5}$$

and whose constituent terms are analytically tractable. Moreover, $\pi(\theta_1 | \eta_1) = N(\theta_1; m_1, K_1)$ and $\pi(\theta_2 | \eta_2) = N(\theta_2; m_2, K_2)$ are multivariate normal densities, $\pi(V^j)$ and $\pi(W^j)$ are the prior densities ascribed to V^j and W^j, $\pi(\eta_1)$, $\pi(\eta_2)$ and $\pi(\eta_3)$ are the prior densities ascribed to η_1, η_2 and η_3.

The marginal likelihood can be efficiently evaluated using a forward filter. It will be helpful here to define

$$\tilde{X}_{t_i} \equiv X_{t_i} - \theta_1 \sin\frac{\pi t_i}{6} - \theta_2 \cos\frac{\pi t_i}{6} = \tilde{F}_{t_i}\theta_{3,t_i} + \nu_{t_i},$$
$$\nu_{t_i} \overset{indep}{\sim} \text{N}(0, \text{diag}\{V^1, \ldots, V^{n_z}\}),$$

so that

$$\tilde{X}_{t_i} | \theta_{3,t_i} \sim N(\tilde{F}_{t_i}\theta_{3,t_i}, \text{diag}\{V\}),$$

where $\tilde{F}_{t_i}$ is the $n_z \times n_z$ identity matrix and will be omitted for ease of notation in what follows. We also write

$$\theta_{3,t_i} | \theta_{3,t_{i-1}} \sim N(\theta_{3,t_{i-1}}, \tilde{W}_{t_i})$$

where $\tilde{W}_{t_i} = k_{t_i}^2(\text{diag}\{W^1, \ldots, W^{n_z}\} + K_3)$.

Algorithm 1 Forward filter

1. Initial distribution: $\theta_{3,t_0} \sim N(m_0, C_0)$. Store the values of m_0 and C_0.
2. For $t_i, i = 1, \ldots, n$,
   (a) Prior at t_i. Using the system equation, we have that
   $$\theta_{3,t_i} | \tilde{x}_{t_{1:i-1}} \sim N(m_{t_{i-1}}, C_{t_{i-1}} + \tilde{W}_{t_i}).$$
   Store $R_{t_i} = C_{t_{i-1}} + \tilde{W}_{t_i}$.
   (b) One step forecast. Using the observation equation, we have that
   $$\tilde{X}_{t_i} | \tilde{x}_{t_{1:i-1}} \sim N(m_{t_{i-1}},\ R_{t_i} + \text{diag}\{V\}).$$
   Store the marginal likelihood contribution
   $$\pi(\tilde{x}_{t_i} | \tilde{x}_{t_{1:i-1}}) = N(\tilde{x}_{t_i}; m_{t_{i-1}}, R_{t_i} + \text{diag}\{V\}).$$
   (c) Posterior at t_i: $\theta_{3,t_i} | \tilde{x}_{t_{1:i}} \sim N(m_{t_i}, C_{t_i})$ where
   $$m_{t_i} = m_{t_{i-1}} + A_{t_i}(\tilde{x}_{t_i} - m_{t_{i-1}}),$$
   $$C_{t_i} = R_{t_i} - A_{t_i} Q_{t_i} A_{t_i}^T,$$
   where $A_{t_i} = R_{t_i} Q_{t_i}^{-1}$ and $Q_{t_i} = R_{t_i} + \text{diag}\{V\}$. Store the values of m_{t_i} and C_{t_i}.

Algorithm 1 gives the steps of the forward filter. We see that the constituent terms in Eq. (5) are obtained from the forward pass as

$$\pi(\tilde{x}_{t_i} | \tilde{x}_{t_{1:i-1}}, \theta_1, \theta_2, \eta_3, V, W) = \text{N}(\tilde{x}_{t_i}; m_{t_{i-1}}, R_{t_i} + V),$$

where $R_{t_i} = C_{t_{i-1}} + \tilde{W}_{t_i}$ and $m_{t_{i-1}}, C_{t_{i-1}}$ are updated recursively; we refer the reader to Petris et al. [4] (see also [3,14,15]) for further details.

Although the marginal likelihood is tractable, the posterior will typically be unavailable in closed form. Hence we use Metropolis–Hastings to generate draws from $\pi(\psi | \tilde{x})$; see Algorithm 2.

Algorithm 2 MCMC scheme

1 Initialise the chain with $\psi^{(0)}$. Set $r = 1$.
2 Propose $\psi^* \sim q(\psi^* | \psi^{(r-1)})$.
3 Calculate the acceptance probability $\alpha(\psi^* | \psi^{(r-1)})$ of the proposed move, where
$$\alpha(\psi^* | \psi^{(r-1)}) = \min\left\{1, A(\psi^* | \psi^{(r-1)})\right\} = \min\left\{1, \frac{\pi(\psi^* | \tilde{x}_{1:n}) q(\psi^{(r-1)} | \psi^*)}{\pi(\psi^{(r-1)} | \tilde{x}_{1:n}) q(\psi^* | \psi^{(r-1)})}\right\}$$
4 With probability $\alpha(\psi^* | \psi^{(r-1)})$, set $\psi^{(r)} = \psi^*$; otherwise set $\psi^{(r)} = \psi^{(r-1)}$.
5 Set $r := r + 1$. Return to step 2.

It remains that, given draws of $\psi^{(1)}, \ldots, \psi^{(N)}$ we can sample $\theta_3^{(r)} \sim \pi(\theta_3 | \psi, x)$, $r = 1, \ldots, N$. This can be achieved by noting the factorisation

$$\pi(\theta_3 | \psi, x) = \pi(\theta_{3,t_n} | \psi, x_{t_{1:n}}) \prod_{i=0}^{n-1} \pi(\theta_{3,t_i} | \theta_{3,t_{i+1}}, \psi, x_{t_{1:i}})$$

where the constituent densities are tractable and can be sampled recursively via a backward sampling algorithm. The key steps are given in Algorithm 3.

Algorithm 3 Backward sampler

3. Sample $\theta_{3,n}|\tilde{x}_{t_{1:n}} \sim N(m_n, C_n)$.
4. For $t_i, i = n, \dots, 1$,

   (a) Backward distribution: $\theta_{3,t_i}|\theta_{3,t_{i+1}}, \tilde{x}_{t_{1:i}} \sim N(h_{t_i}, H_{t_i})$, where

$$h_{t_i} = m_{t_i} + C_{t_i}(C_{t_i} + \tilde{W}_{t_{i+1}})^{-1}(\theta_{3,t_{i+1}} - m_{t_i}),$$
$$H_{t_i} = C_{t_i} - C_{t_i}(C_{t_i} + \tilde{W}_{t_{i+1}})^{-1}C_{t_i}.$$

   (b) Sample $\theta_{3,t_i}|\theta_{3,t_{i+1}}, \tilde{x}_{t_{1:i}} \sim N(h_{t_i}, H_{t_i})$.

Missing data

Missing observations are commonplace, that is, only observations on a subset of components of X_t may be available at time t_i. To account for this in the model we let $\tilde{X}^o_{t_i}$ denote the observed rates at time t_i. The observation model is then written as

$$\tilde{X}^o_{t_i} = P_{t_i}\tilde{X}_{t_i} \tag{6}$$

where the $n_{obs} \times n_z$ incidence matrix P_{t_i} determines which components are observed at time t_i [12]. For example, if we have data from 5 zones and data are missing at the second and third zone at time t_i, then the incidence matrix is

$$P_{t_i} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The forward filter and backward sampler can be modified straightforwardly to allow for this scenario. In brief, each occurrence of $\tilde{F}_{t_i}$ is replaced by $P_{t_i}\tilde{F}_{t_i}$ and each occurrence of V is replaced by $P_{t_i}VP^T_{t_i}$ in Algorithm 1, which extracts the relevant portion of the variance–covariance matrix corresponding to the observed data points.

4.2. Within-sample predictive density

In order to assess model fit, we consider the within-sample predictive density. The within-sample predictive density is given by

$$\pi(\hat{x}_{t_{1:n}}|x_{t_{1:n}}) = \iint \pi(\hat{x}_{t_{1:n}}|\theta_{3,t_{1:n}}, \psi)\pi(\theta_{3,t_{1:n}}, \psi|x_{t_{1:n}})d\theta_{3,t_{1:n}}d\psi$$

where

$$\pi(\theta_{3,t_{1:n}}, \psi|x_{t_{1:n}}) = \pi(\theta_{3,t_{1:n}}|\psi, x_{t_{1:n}})\pi(\psi|x_{t_{1:n}}).$$

Although the within-sample predictive density is intractable, draws from $\pi(\theta_{3,t_{1:n}}, \psi|x_{t_{1:n}})$ are readily available and therefore $\pi(\hat{x}_{t_{1:n}}|x_{t_{1:n}})$ can be obtained via Monte Carlo. Given draws $(\psi^{(r)}, \theta^{(r)}_{3,t_{1:n}}), r = 1, \dots, N$, we can simulate

$$\hat{X}^{(r),j}_{t_i}|\theta^{(r),j}_{t_i}, \psi^{(r),j} \sim N(F_{t_i}\theta^{(r),j}_{t_i}, V^{(r),j}), \qquad r = 1, \dots, N, \quad i = 1, \dots, n, \quad j = 1, \dots, n_z, \tag{7}$$

where $\theta^{(r),j}_{t_i} = (\theta^{(r),j}_1, \theta^{(r),j}_2, \theta^{(r),j}_{3,t_i})$ denotes the rth sample of $\theta^j_{t_i}$, with $\hat{X}^{(r),j}_{t_i}$ defined similarly. Draws obtained from (7) can be summarised (e.g. via the mean, upper and lower quantiles) and bench-marked against the observed data.

4.3. k-Step ahead prediction

The system and observation forecast distributions can be obtained by exploiting the linear Gaussian structure of the DLM. The one-step ahead system forecast density is given by

$$\pi(\theta_{3,t_{n+1}}|x_{t_{1:n}}) = \iint \pi(\theta_{3,t_{n+1}}|\theta_{3,t_n}, \psi, x_{t_{1:n}})\pi(\theta_{3,t_n}|\psi, x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\theta_{3,t_n}d\psi$$
$$= \int \pi(\theta_{3,t_{n+1}}|\psi, x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\psi$$

where

$$\pi(\theta_{3,t_{n+1}}|\psi, x_{t_{1:n}}) = N(\theta_{3,t_{n+1}}; m_{t_n}, C_{t_n} + \tilde{W}_{t_{n+1}}).$$

Similarly, the one-step ahead observation forecast density is given by

$$\pi(x_{t_{n+1}}|x_{t_{1:n}}) = \int \pi(x_{t_{n+1}}|\psi, x_{t_{1:n}})\pi(\psi|x_{t_{1:n}})d\psi$$

where

$$\pi(x_{t_{n+1}}|\psi, x_{t_{1:n}}) = N(x_{t_{n+1}}; m_{t_n}, C_{t_n} + \tilde{W}_{t_{n+1}} + V).$$

Hence, given N posterior summaries $(m^{(r)}_{t_n}, C^{(r)}_{t_n}), r = 1, \dots, N$ from $\pi(\theta_{3,t_n}|\psi, x_{t_{1:n}})$ and $\psi^{(r)}$ from $\pi(\psi|x_{t_{1:n}})$, the one-step ahead state and observation forecast distributions can be sampled via Monte Carlo, by drawing

$$\theta^{(r)}_{3,t_{n+1}}|\psi^{(r)}, x_{t_{1:n}} \sim N(m^{(r)}_{t_n}, C^{(r)}_{t_n} + \tilde{W}^{(r)}_{t_{n+1}}),$$
$$\tilde{X}^{(r)}_{t_{n+1}}|\psi^{(r)}, x_{t_{1:n}} \sim N(m^{(r)}_{t_n}, C^{(r)}_{t_n} + \tilde{W}^{(r)}_{t_{n+1}} + V^{(r)}).$$

Then, $X^{(r)}_{t_{n+1}}$ can be obtained from $\tilde{X}^{(r)}_{t_{n+1}}$ by adding the term $\theta^{(r)}_1 \sin\frac{\pi t_{n+1}}{6} + \theta^{(r)}_2 \cos\frac{\pi t_{n+1}}{6}$ to the latter. For the general k-step ahead forecast, the above draws are replaced by

$$\theta^{(r)}_{n+k}|\psi, x_{t_{1:n}} \sim N\left\{m^{(r)}_{t_n}, R^{(r)}_{t_{n+k}}\right\},$$
$$\tilde{X}^{(r)}_{t_{n+k}}|\psi^{(r)}, x_{t_{1:n}} \sim N\left\{m^{(r)}_{t_n}, R^{(r)}_{t_{n+k}} + V^{(r)}\right\},$$

where

$$R^{(r)}_{t_{n+k}} = C^{(r)}_{t_n} + \sum_{i=1}^{k} \tilde{W}^{(r)}_{t_{n+i}}.$$

5. Application

In what follows, and where required, we implement the MCMC scheme from Section 4 by taking a random walk proposal with Gaussian innovations. We have that $q(\psi^*|\psi) = N(\psi^*; \psi, \Sigma)$ where the innovation matrix $\Sigma = \gamma\widehat{Var}(\psi|x)$, with $\widehat{Var}(\psi|x)$ obtained from a pilot run and γ is chosen to give an acceptance rate of around 25% [18]. Within the MCMC scheme, for mathematical convenience, we will work with precisions so that $\tau_V = 1/V, \tau_W = 1/W$. Moreover, for parameter vectors whose components must be strictly positive (i.e. V, W, η) we implement the proposal on the log scale.

The inference scheme was coded in R [19]; code to reproduce the analysis can be found at https://doi.org/10.25405/data.ncl.24204726.v1.

5.1. Single zone analysis

In this section we assess the assumption that amplitude and phase vary with time. We present results for zone 4 and note similar findings (namely that amplitude and phase are plausibly constant) for the remaining zones.

For the single zone model, $\psi = (\tau_V, \tau_{W_1}, \tau_{W_2}, \tau_{W_3})'$ is the vector of precision parameters. We set the mean and variance of θ_{t_0} to be $m_0 = (1.5, 1.5, 6)$ and $C_0 = \text{diag}\{1.5, 1.5, 20\}$ respectively. We take an uninformative and independent prior specification for the components of ψ, via $\tau_V, \tau_{W_1}, \tau_{W_2}, \tau_{W_3} \sim Ga(0.1, 0.1)$. We assessed convergence of the MCMC simulations using Geweke's diagnostic, the Gelman–Rubin statistic, and effective sample size (ESS) across six distinct chains, confirming robust convergence and sampling efficiency for all parameters. Each MCMC run used 22k iterations with the first 2k iterations discarded as burn-in, leaving 20k iterations on which to base posterior summaries.

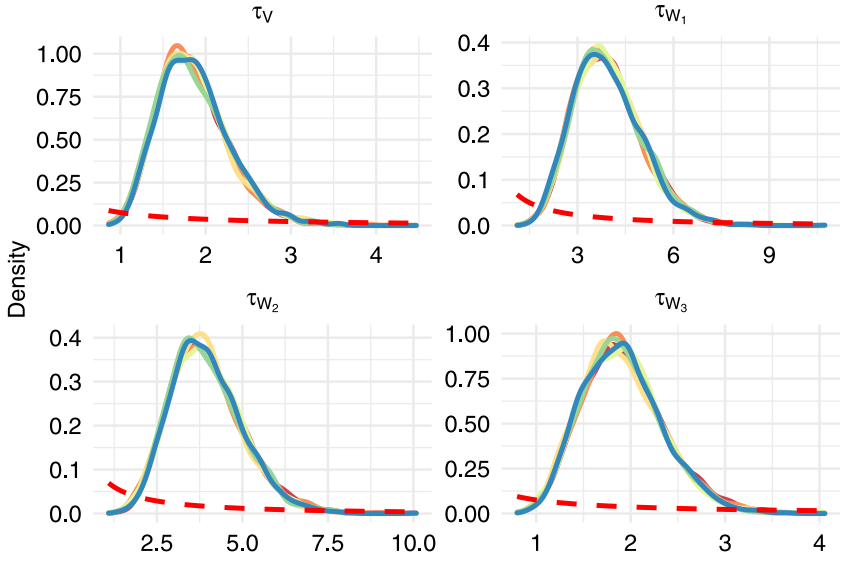

Fig. 4. Density plots of $\tau_V, \tau_{W_1}, \tau_{W_2}, \tau_{W_3}$ respectively, from multiple chains of 20k iterations and a thin of 20 with prior densities overlaid in red.

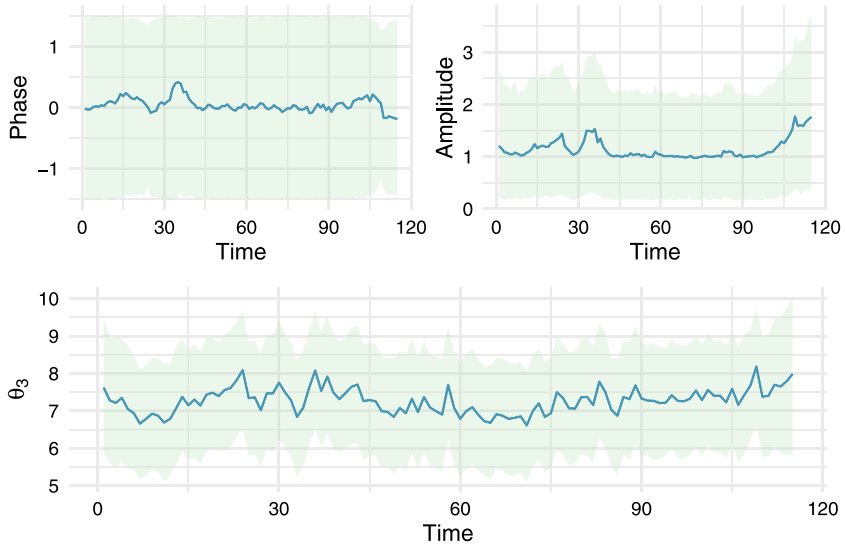

Fig. 5. Phase, amplitude and θ_{3,t_i} mean and 95% CI at zone 4 from time t_1 to t_{115}.

The marginal MH scheme gives the estimated marginal posterior densities for the components of ψ shown in Fig. 4 with their prior densities overlaid. The ψ samples were thinned to obtain 1k (near uncorrelated) draws form the marginal parameter posterior, denoted $\{\psi^{(r)}\}_{r=1}^{1000}$. The FFBS algorithm was then executed for each $\psi^{(r)}$, to obtain samples of the dynamic parameter vector, $\{\theta_{t_i}^{(r)}\}_{r=1}^{1000}$, $i = 1, \ldots, n$, from the within-sample predictive. Samples of the dynamic components θ_{1,t_i} and θ_{2,t_i} can be transformed via (3) to obtain phase and amplitude draws from their respective within-sample predictive densities; see Section 4.2 for further details regarding the method for obtaining samples from these predictive distributions. These distributions are summarised in Fig. 5 via their means and 95% credible intervals. We can conclude that, upon allowing for the uncertainty in amplitude and phase, they are plausibly constant over time for this zone. Performing the analysis on the remaining zones shows that the same conclusions can be drawn. This suggests that the dynamic parameters θ_{1,t_i} and θ_{2,t_i}, $i = 1, \ldots, n$, can reasonably be replaced with static parameters θ_1 and θ_2.

We assess the validity of the proposed model for a single zone by comparing observed data with their model-based within-sample posterior predictive distributions and with model-based out-of-sample forecast distributions. For the latter, we withheld the last 10 observations when fitting the model. Fig. 6 shows the within-sample predictive distribution for the observation process, summarised by the mean and 95% credible interval calculated for each time point. This suggests that the model is able to reasonably account for the observation process. Similarly, the 10-step ahead forecast distribution is summarised by the mean and 95% credible interval at each time point. We see that the forecast distribution is able to capture the general trend exhibited by the observations.

5.2. Joint zone analysis

We now consider the joint model over all zones detailed in Section 3.2. Our prior specification takes the following form.

We expect that amplitude and phase should be similar at nearby zones. Recall that $\theta_1 \sim GP(m_1(\cdot), f_1(\cdot;\eta_1))$, $\theta_2 \sim GP(m_2(\cdot), f_2(\cdot;\eta_2))$ and the Gaussian process components in the dynamic mean process are $p_{t_i} \overset{indep}{\sim} GP\{0, f_3(\cdot;\eta_3)\}$. We take the mean functions to be constant so that $m_1(\cdot) = m_2(\cdot) = 1.5\mathbf{1}$, with $\mathbf{1}$ defined as an $n_z \times 1$ vector of 1 s. We have that $f_k(d_{jj'};\eta_k) = \sigma_k^2 \exp(-\phi_k d_{jj'})$, $k = 1, 2, 3$. We take

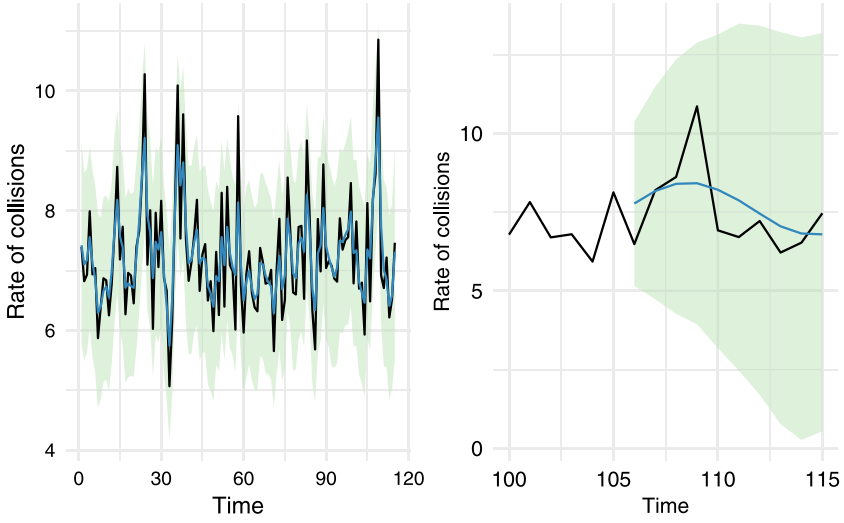

Fig. 6. Left: Zone 4 observed data (black) with overlaid within-sample predictions — mean (blue) and 95% credible intervals (green). Right: Observed data (black) with overlaid 10 step ahead predictions — mean (blue) with 95% credible interval (green).

$\log \sigma_k \overset{indep}{\sim} N(\log(0.1), 0.1)$ representing fairly strong prior beliefs about the amplitude variance and phase within a zone. For the logarithm of the inverse length scales, we take $\log \phi_k \sim N(\log(0.1), 0.1)$ giving typical length scales of around 10 km, reflecting typical distances between zones. The precisions of the observation equations governing each zone are $\tau_{V^j} \overset{indep}{\sim} Ga(0.1, 0.1)$ and similarly for the system variances, $\tau_{W^j} \overset{indep}{\sim} Ga(0.1, 0.1)$, $j = 1, \ldots, n_z$. Finally, the initial values θ^j_{3,t_0} for each site were assumed to follow $N(6, 20)$ distributions.

The MCMC scheme was run for 10^6 iterations; the output is summarised by Table 1. Fig. 7 shows the mean value and 95% credible interval of the posterior densities for amplitude and phase at each zone against longitude. There are signs of spatial dependence as the phase seems to decrease and amplitude increases in zones further to the east. Fig. 8 shows a single period of the sine curve, averaged over draws of amplitude and phase for the most eastern versus most western zone with 95% credible intervals. From this we would expect to see more pronounced fluctuations in the rate of collisions across the year for eastern zones. Furthermore, we would expect the highest rate of collisions to be a month sooner (August) in eastern zones than that in western zones (September).

Fig. 9 shows summaries (mean and 95% credible interval) of the difference between observations and the within-sample predicted observation process for zones 2, 4 and 8. The left-hand-side plots show the differences from the single zone analysis and the right-hand-side from the joint analysis. It is clear that the mean difference at each time-zone combination is small and that a mean difference of zero is plausible (the 95% credible intervals include zero). Comparing left to right, shows the improvement in the within-sample predictions from a single zone analysis to a joint model; that is, the spatial information included through the GP has increased prediction precision. We additionally calculated the root mean square error (RMSE) at each time-point (observation vs. prediction) and averaged this measure over all time points for each zone; the results are shown in Table 2. We see that the mean RMSEs are approximately 5 times larger for the single zone analysis, giving further evidence of an improvement in fit when considering a joint model over all zones.

Fig. 10 shows 10-step ahead predictions for zones 2, 4, 6 and 8, following application of the method in Section 4.3. Note that the last 10 observations were removed from each zone before running the inference scheme. The figure shows that the forecast distributions are consistent with the data as they lie within the forecast intervals for all zones. As we would expect, uncertainty grows as we move away from the last recorded observation.

Table 1
Marginal parameter posterior means and quantile-based 95% credible intervals obtained from the MCMC scheme.

ψ	Mean	95% CI	ψ	Mean	95% CI
V^1	0.034	(0.021, 0.052)	θ_1^4	0.251	(−0.045, 0.532)
V^2	0.025	(0.015, 0.039)	θ_1^5	0.226	(−0.065, 0.514)
V^3	0.059	(0.039, 0.084)	θ_1^6	0.249	(−0.039, 0.526)
V^4	0.037	(0.022, 0.058)	θ_1^7	−0.181	(−0.501, 0.133)
V^5	0.031	(0.018, 0.048)	θ_1^8	−0.014	(−0.308, 0.271)
V^6	0.041	(0.023, 0.066)	θ_2^1	0.585	(0.301, 0.877)
V^7	0.119	(0.059, 0.196)	θ_2^2	0.651	(0.367, 0.944)
V^8	0.045	(0.026, 0.071)	θ_2^3	0.566	(0.285, 0.856)
W^1	0.021	(0.011, 0.037)	θ_2^4	0.424	(0.144, 0.722)
W^2	0.024	(0.012, 0.041)	θ_2^5	0.809	(0.530, 1.098)
W^3	0.023	(0.011, 0.044)	θ_2^6	0.601	(0.311, 0.896)
W^4	0.025	(0.012, 0.044)	θ_2^7	1.264	(0.931, 1.587)
W^5	0.024	(0.013, 0.043)	θ_2^8	0.945	(0.660, 1.249)
W^6	0.034	(0.016, 0.061)	σ_1	1.688	(1.315, 2.309)
W^7	0.099	(0.031, 0.213)	σ_2	1.545	(1.201, 2.253)
W^8	0.029	(0.014, 0.055)	σ_3	1.352	(1.349, 1.355)
θ_1^1	0.357	(0.066, 0.642)	ϕ_1	1.527	(1.278, 1.917)
θ_1^2	0.213	(−0.077, 0.494)	ϕ_2	1.603	(1.387, 1.903)
θ_1^3	0.213	(−0.084, 0.499)	ϕ_3	1.103	(1.098, 1.107)

Table 2
The mean RMSE over all time-points for each zone, from the single zone and joint zone analyses.

Zone	Mean RMSE	
	Single zone	Joint zone
1	1.384	0.198
2	1.066	0.167
3	1.214	0.258
4	1.083	0.206
5	1.114	0.185
6	1.191	0.217
7	1.339	0.378
8	1.107	0.227

6. Discussion and limitations

We have developed a spatio-temporal model for collision rates that allows for serial dependence, seasonality and correlation between rates at nearby zones. We considered a dynamic linear model (DLM)

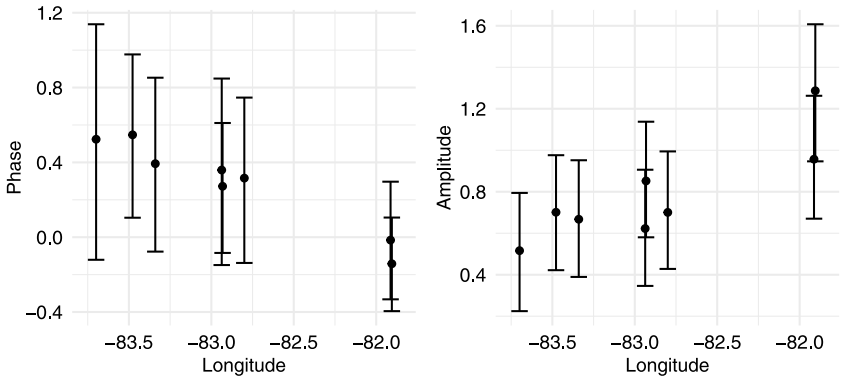

Fig. 7. Mean amplitude and phase with 95% credible intervals against longitude for each zone.

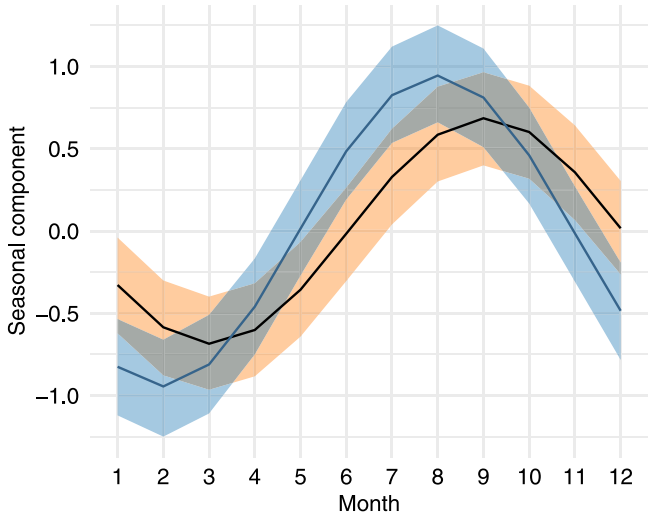

Fig. 8. Mean and 95% credible intervals for the seasonal component for the most western zone (black) against the most eastern (blue).

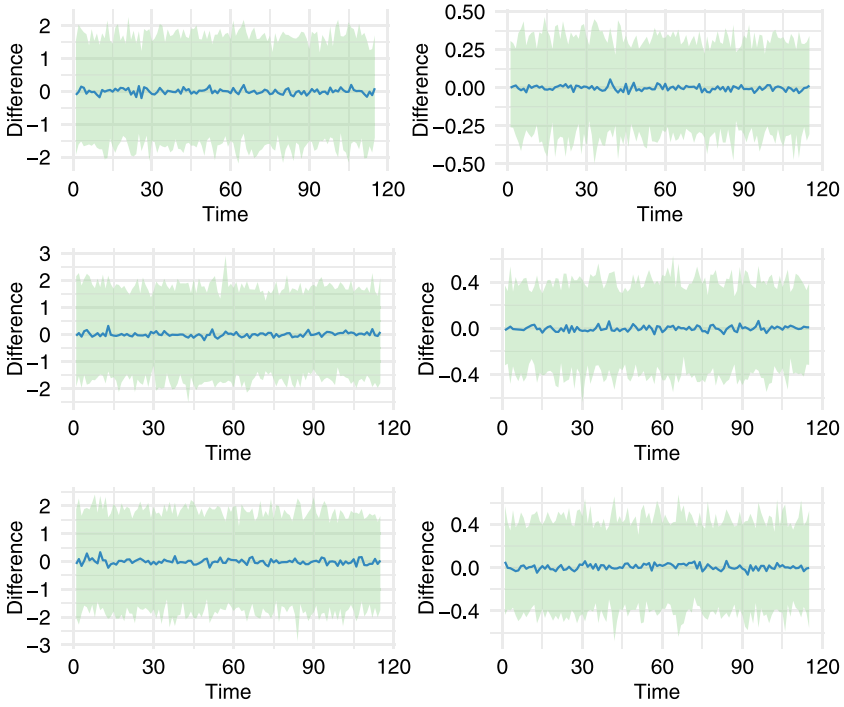

Fig. 9. Mean (blue) and 95% credible intervals (green) for the difference between the within-sample predictive and the observations over time. Each row shows the differences from the single zone analysis (left) and the joint zone analysis (right) for zones 2 (top), 4 (middle) and 8 (bottom).

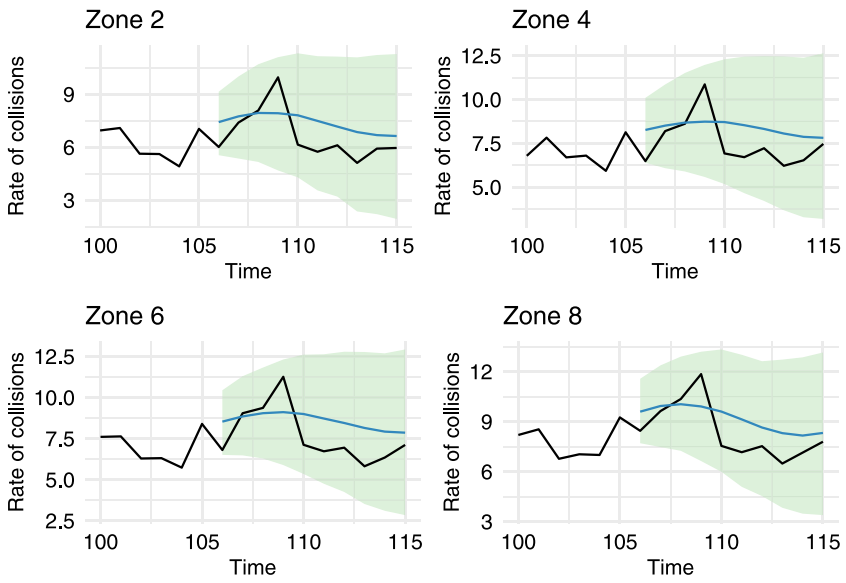

Fig. 10. Rates of collisions in Florida zones 2, 4, 6 and 8 with overlaid out-of-sample 10-step ahead predictions — mean (blue) and 95% credible intervals (green).

whose observation equation takes the form of a single harmonic with a smoothly time-varying amplitude and phase, thus accounting for seasonality and potential long term changes. Spatial consistency is accounted for at nearby zones by adding a Gaussian process (GP) component in the system equation. The model can be fitted in Bayesian paradigm using an efficient two-stage Markov chain Monte Carlo procedure, targeting the joint posterior over the parameters, the latent time-varying harmonic coefficients (amplitude and phase) and dynamic mean. At the first stage, parameter samples are generated from the marginal parameter posterior using a random walk Metropolis algorithm with the likelihood evaluated via a forward filter. At the second stage, samples of the dynamic parameters are generated conditionally on the static parameter draws from stage one using a backward sampler. Further details of this FFBS approach can be found in [4] (see also [14,15]).

We applied our approach to a data set consisting of 115 months of collision rates over eight Traffic Administration Zones in Florida and Georgia, USA. An exploratory analysis that considered separate models for each zone found that the phase and amplitude were plausibly constant. We were therefore able to simplify the joint model over all zones by treating the harmonic components as static, with a GP prior allowing correlation between these parameters at nearby zones. The validity of both the single zone and joint models was assessed using within-sample posterior predictive distributions, which suggested a satisfactory fit in both cases. Moreover, the within-sample predictions were improved substantially when using the joint model, with the credible intervals of our predictions narrowing almost fivefold, and a reduction in root mean squared error (RMSE) between the observations and predictions of around a factor of 5.

Our analysis suggests clear spatial patterns between phase and longitude and amplitude and longitude. For all zones we found that the lowest rates of collisions would fall earlier in the year. The model also suggests that for western zones, the lowest rates would be in March, and in February for eastern zones. It appears that peak collision rates are in September in the East and August in the West. We would also expect to see a larger fluctuation in the rate of collisions in an eastern zone. Our interest also lies in the ability to forecast collision rates in future months. Model-based out-of-sample forecast distributions suggest that our model is able to capture observed trend and seasonality in monthly collision rates up to around a year ahead.

Our modelling approach can be improved in a number of ways. For example, it is common to have covariate information such as traffic flow or average speed associated with a particular location at which a collision has occurred. However, pooling such data over zones is time-consuming and not always straightforward. Nevertheless, incorporation of covariates into the DLM framework is straightforward in principle, via the observation equation, and we anticipate improved prediction in this scenario. Although not pursued here, our model can also be used to predict collision rates at zones for which observations are not available. Interpolation of the fitted GP component in the system equation governing the dynamic mean and GP prior over the static parameters governing the harmonic, can be performed for unobserved zones of interest; see e.g. [20] for further details. In scenarios with many zones, it may be beneficial to model spatial dependence via a conditional autoregressive (CAR) process, which can be seen as a special case of our approach (see e.g. [21]). In particular, conditioning the latent dynamic component in the DLM via a local neighbourhood structure could lead to a more computationally efficient inference scheme.

CRediT authorship contribution statement

Nicola Hewett: Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Andrew Golightly:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Lee Fawcett:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Neil Thorpe:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix. Additional convergence checks

We assessed the convergence of the MCMC simulations using multiple diagnostics to ensure the reliability of parameter estimates. Convergence diagnostics included Geweke's diagnostic, the potential scale

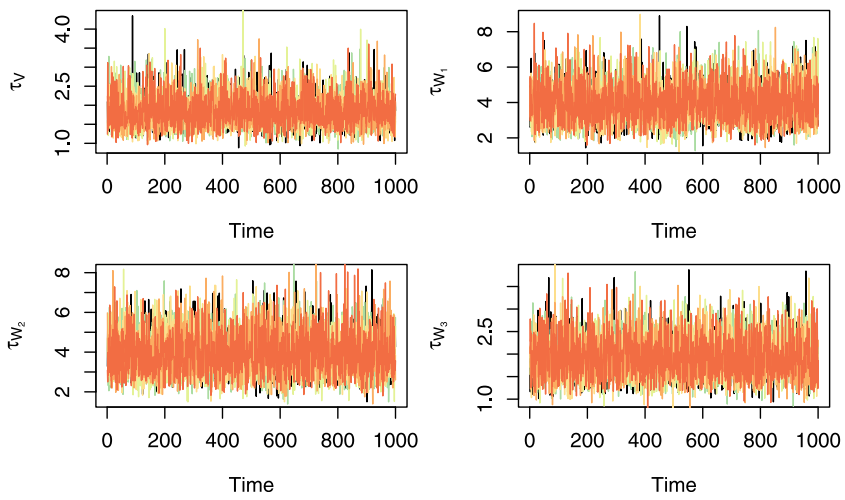

Fig. 11. Trace plots for ψ from the single zone analysis across multiple MCMC chains. Each line represents one of the six independent chains used in the analysis.

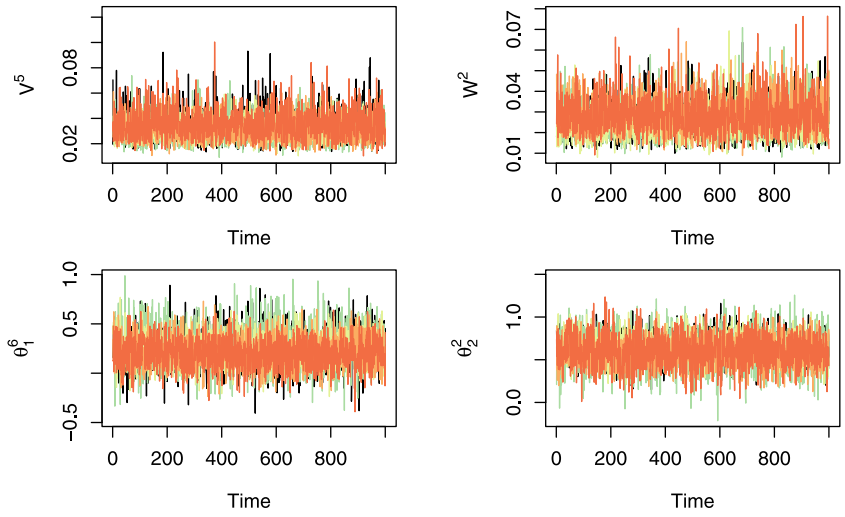

Fig. 12. Trace plots for a subset of ψ from the joint zone analysis across multiple MCMC chains. Each line represents one of the six independent chains used in the analysis.

reduction factor (PSRF) as per Gelman–Rubin, and the Effective Sample Size (ESS) (see e.g. [22]). These diagnostics were obtained using the R `coda` package [23]. Geweke's diagnostic was applied to evaluate the equality of the means of the first 10% and the last 50% of the chains, across all parameters and six independent runs. The diagnostic yielded z-scores within the range of -1.3785 to 1.8647, suggesting no significant difference between the early and late phases of the chains, thus indicating convergence of all chains and parameters. The Gelman–Rubin PSRF was calculated for each parameter, with all point estimates and upper confidence intervals equalling 1. This indicates that between-chain variance is negligible compared to within-chain variance, further supporting convergence across all parameter chains. ESS was computed for each parameter, revealing sizes of 6000 for V and W_3, 6189.358 for W_1, and 6773.519 for W_2. These suggest adequate sampling efficiency, providing a solid basis for the posterior estimates derived from the chains. Together, these diagnostics provide strong evidence of convergence and adequate sampling within the MCMC simulations, supporting the robustness of the parameter estimates obtained. Fig. 11 shows trace plots for the parameters $\psi = (\tau_V, \tau_{W_1}, \tau_{W_2}, \tau_{W_3})$, derived from six distinct chains. The overlapping and mixing of chains suggest good convergence behaviour for each parameter, indicating that they have reached their stationary distributions. Similar tests were performed for the joint analysis which also showed good convergence for all parameters (example trace plots shown in Fig. 12).

References

[1] WHO, Road traffic injuries, 2020, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. (Online; Accessed 24 May 2021).
[2] A. Harvey, Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge University Press, 1990.
[3] M. West, J. Harrison, Bayesian Forecasting and Dynamic Models, Springer Science & Business Media, 2006.
[4] G. Petris, S. Petrone, P. Campagnoli, Dynamic linear models, in: Dynamic Linear Models with R, Springer, 2009, pp. 31–84.
[5] D. Gamerman, H.S. Migon, Dynamic hierarchical models, J. R. Stat. Soc. Ser. B Stat. Methodol. 55 (3) (1993) 629–642.
[6] X. Fei, C.-C. Lu, K. Liu, A Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction, Transp. Res. C 19 (6) (2011) 1306–1318.
[7] A. Boulieri, S. Liverani, K. Hoogh, M. Blangiardo, A space–time multivariate Bayesian model to analyse road traffic accidents by severity, J. R. Statist. Soc. Ser. A 180 (1) (2016) 119–139.
[8] X. Xiong, L. Chen, J. Liang, A new framework of vehicle collision prediction by combining SVM and HMM, IEEE Trans. Intell. Transp. Syst. 19 (3) (2018) 699–710.
[9] C. Dong, K. Xie, X. Sun, M. Lyu, H. Yue, Roadway traffic crash prediction using a state-space model based support vector regression approach, PLoS One 14 (4) (2019) e0214866.
[10] V. Popov, G. Nightingale, A.J. Williams, P. Kelly, R. Jepson, K. Milton, M. Kelly, Trend shifts in road traffic collisions: An application of hidden Markov

models and generalised additive models to assess the impact of the 20 mph speed limit policy in Edinburgh, Environ. Plan. B: Urban Anal. City Sci. 48 (9) (2021) 2590–2606.
[11] P.N.V.S.R. Buddhavarapu, On Bayesian Estimation of Spatial and Dynamic Count Models Using Data Augmentation Techniques: Application to Road Safety Management (Ph.D. thesis), 2015.
[12] Y. Lai, A. Golightly, R.J. Boys, Sequential Bayesian inference for spatio-temporal models of temperature and humidity data, J. Comput. Sci. 43 (2020) 101125.
[13] G. Shaddick, J. Wakefield, Modelling daily multivariate pollutant data at multiple sites, J. R. Stat. Soc. Ser. C. Appl. Stat. 51 (3) (2002) 351–372.
[14] C.K. Carter, R. Kohn, On Gibbs sampling for state space models, Biometrika 81 (3) (1994) 541–553.
[15] S. Frühwirth-Schnatter, Data augmentation and dynamic linear models, J. Time Ser. Anal. 15 (2) (1994) 183–202.
[16] S. Banerjee, B.P. Carlin, A.E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, CRC Press, 2014.
[17] W.R. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte Carlo in Practice, CRC Press, 1995.
[18] G.O. Roberts, J.S. Rosenthal, Optimal scaling for various Metropolis–Hastings algorithms, Statist. Sci. 16 (4) (2001) 351–367.
[19] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021.
[20] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2005.
[21] A.N. Pettitt, I.S. Weir, A.G. Hart, A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data, Stat. Comput. 12 (2001) 353–367.
[22] D. Gamerman, H. Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition, second ed., Chapman and Hall/CRC, 2006, http://dx.doi.org/10.1201/9781482296426.
[23] M. Plummer, N. Best, K. Cowles, K. Vines, CODA: convergence diagnosis and output analysis for MCMC, R News 6 (1) (2006) 7–11.

Nicola Hewett is a Lecturer of Mathematical Sciences at Newcastle University, UK. She earned her Ph.D. from the same university in 2023 and her research focuses on applying statistical analysis to improve road safety.

Andrew Golightly is Professor of Statistics in the Department of Mathematical Sciences at Durham University. He received his Ph.D. from Newcastle University in 2006 and was appointed to a Lectureship in the same year. He was promoted to Senior Lecturer in 2016 and Reader in 2019 before moving to Durham University in 2022 and taking up a chair in 2023. Andrew is particularly interested in computational Bayesian inference and regularly publishes in this area.

Lee Fawcett is Reader in Applied Statistics in the School of Mathematics, Statistics and Physics, and Associate Dean (Global), at Newcastle University in the UK. Lee's primary research areas include environmental extremes and statistical methods for road safety. He also has a passion for Global Education initiatives, including setting up mobility opportunities for under-represented student groups.

Neil Thorpe is a Senior Associate Director at Jacobs in the UK currently engaged in a range of transport planning projects in the UK and overseas. Prior to joining Jacobs in 2022, Neil was a Senior Lecturer in Transport and Head of the Future Mobility Group in the School of Civil Engineering and Geosciences at Newcastle University, from where he also received his Ph.D. in 2005. His main research focus at Newcastle University was the analysis of road traffic collision data to inform the evaluation of road safety countermeasures and to predict the location of road traffic collision hotspots.