INTERFACE

Research

Authors for correspondence:
Elizaveta Semenova
e-mail: elizaveta.p.semenova@gmail.com
Seth Flaxman
e-mail: seth.flaxman@cs.ox.ac.uk

†Co-first authors.

THE ROYAL SOCIETY
PUBLISHING

PriorVAE: encoding spatial priors with variational autoencoders for small-area estimation

Elizaveta Semenova[1,†], Yidan Xu[2,†], Adam Howes[3], Theo Rashid[3], Samir Bhatt[3,4], Swapnil Mishra[3,4] and Seth Flaxman[1]

[1]University of Oxford, Oxford, UK
[2]University of Michigan, Ann Arbor, MI, USA
[3]Imperial College London, London, UK
[4]University of Copenhagen, Kobenhavn, Denmark

ES, 0000-0002-8271-2575; AH, 0000-0003-2386-4031; SB, 0000-0002-0891-4611; SM, 0000-0002-8759-5902; SF, 0000-0002-2477-4217

Gaussian processes (GPs), implemented through multivariate Gaussian distributions for a finite collection of data, are the most popular approach in small-area spatial statistical modelling. In this context, they are used to encode correlation structures over space and can generalize well in interpolation tasks. Despite their flexibility, off-the-shelf GPs present serious computational challenges which limit their scalability and practical usefulness in applied settings. Here, we propose a novel, deep generative modelling approach to tackle this challenge, termed PriorVAE: for a particular spatial setting, we approximate a class of GP priors through prior sampling and subsequent fitting of a variational autoencoder (VAE). Given a trained VAE, the resultant decoder allows spatial inference to become incredibly efficient due to the low dimensional, independently distributed latent Gaussian space representation of the VAE. Once trained, inference using the VAE decoder replaces the GP within a Bayesian sampling framework. This approach provides tractable and easy-to-implement means of approximately encoding spatial priors and facilitates efficient statistical inference. We demonstrate the utility of our VAE two-stage approach on Bayesian, small-area estimation tasks.

1. Introduction

Spatially referenced data come in a variety of forms, including exact geographical coordinates such as a latitude and longitude or predefined geographical areal units such as a village, administrative unit or pixel of a raster image. The latter are known as areal unit data, and are found in fields such as epidemiology, environmental and political science; a variety of relevant methods come under the banner of small-area statistics [1]. There are many motivations for modelling such data, from surveillance programme evaluation to identifying environmental risk factors for disease. Small-area statistics are particularly relevant to informing policy decisions, which are often made at the areal unit level [2].

Statistical modelling of spatial data is routinely performed using Gaussian process (GP) priors [3]. GPs have gained popularity in a variety of applied fields due to their flexibility, ease of implementation, and their inherent ability to characterize uncertainty. However, GPs also present a number of practical challenges. For example, basic inference and prediction using a GP require matrix inversions and determinants—both of which scale cubically with data size. This makes applications of GPs prohibitive for large datasets. Moreover, kernel design of a GP requires substantial domain knowledge in order to reflect characteristics of the process of interest [4]. Hence, the choice of an inference

method is of great importance when it comes to dealing with GPs. The theoretic asymptotic convergence properties and diversity of Markov chain Monte Carlo (MCMC) approaches make it the most reliable method for Bayesian inference. However, MCMC scales poorly, and struggles to deal with the high degree of auto-correlation inherent to spatial data, limiting its utility in large spatial settings [5]. A range of more tractable approximate methods have been developed, such as Laplace approximation, variational inference [6], expectation propagation [7], or inducing variables in sparse GPs [8]. However, few of these approximate methods have asymptomatic guarantees or yield consistently accurate posterior estimates [9] over a wide range of challenging datasets. User-friendly software packages, such as R-INLA [10], provide extremely convenient interfaces to a large set of predefined models, but do not provide enough flexibility for custom model development, and hence have limitations for specific classes of applied research. There is an unmet need for approaches that can be easily implemented and customized in popular probabilistic programming languages such as Stan [11], NumPyro [12], PyMC3 [13] or Turing.jl [14], while still scaling favourably to large datasets. Here, we propose a novel computational technique which leverages the variational autoencoder (VAE) model from deep learning and combines it with Bayesian inference [15,16] with the goal of small-area estimation.

An autoencoder [17] is a neural network architecture used for the task of *supervised* dimensionality reduction and representation learning. It is comprised of three components: an encoder network, a decoder network and a low dimensional layer containing a latent representation. The first component, the encoder $\mathcal{E}_\gamma(\cdot)$ with parameters γ, maps an input $x \in \mathcal{R}^p$ into the latent variable $z \in \mathcal{R}^d$, where d is typically lower than p. The decoder network $\mathcal{D}_\psi(\cdot)$ with parameters ψ aims to reconstruct the original data x from the latent representation z. Therefore, an autoencoder imposes a 'bottleneck' layer in the network which enforces a compressed representation z of the original input x. This type of dimensionality reduction technique is particularly useful when some structure exists in the data, such as spatial correlations. This structure can be learnt and consequently leveraged when forcing the input through the network's bottleneck. Parameters of the encoder and decoder networks are learnt through the minimization of a reconstruction loss function $p(x|\hat{x})$ expressed as the likelihood of the observed data x given the reconstructed data $\hat{x}$.

A VAE extends this formulation to a probabilistic generative model [18]. Rather than learning a latent representation z, it learns a probabilistic distribution $p(z|x)$, from which a latent representation can be sampled. This distribution is chosen in such a way that the latent variables z are independent and normally distributed. To achieve this, the encoder outputs a pair of d-dimensional vectors $\mu_\gamma, \sigma_\gamma^2$ to characterize the mean and variance of z, respectively. The latent variable then follows the Gaussian distribution $z \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2 I)$. The loss function is modified to include a Kullback–Leibler (KL) divergence term implied by a standard $\mathcal{N}(0, I)$ prior on z: $\mathcal{L}(x, \hat{x}) = p(x|\hat{x}) + \mathrm{KL}(\mathcal{N}(\mu_\gamma, \sigma_\gamma^2 I)\|\mathcal{N}(0, I))$. The KL term can be interpreted as a regularizer ensuring parsimony in the latent space. New data can be generated by sampling from the latent space with the trained decoder network. This feature of VAEs has led to a series of successful applications with the goal of generating new samples from the approximated distribution of real data [19].

In this paper, we propose a novel use of VAEs, termed PriorVAE: we learn uncorrelated representations of spatial GP priors for a predefined spatial setting, and then use the trained decoder to perform Bayesian inference on new data.

Our contributions can be summarized as follows.

— We introduce a two-stage process for inference. First, we train a VAE to create an uncorrelated representation of complex spatial priors. Next, we use the learnt latent distribution in the model instead of the GP prior to perform MCMC inference on new data, while keeping the trained decoder fixed.
— We demonstrate the usage of VAE priors on a range of simulated and real-life datasets to show their performance in small-area estimation tasks.

The rest of this paper is organized as follows. In §2, we lay out the methodology of the two-stage approach. In §3, we demonstrate the approach on synthetic and real datasets. In §4, we conclude with a discussion and provide a broader outlook on the potential impact of this work.

2. Methods

2.1. Three types of spatial data

There are three types of spatial data: areal (or lattice), geostatistical (or point references) and point patterns. Models of all three of them rely on the notion of GPs and their evaluations in the form of the multivariate normal distribution. Areal data arise when a fixed domain is partitioned into a finite number of sub-regions at which outcomes are aggregated. Number of disease cases recorded per county, regardless of the exact locations where within the county the cases occurred or got recorded, constitutes an example of areal data. State-of-the-art models of areal data rely on 'borrowing strength' from neighbours and use a hierarchical Bayesian formulation to do so. Widely used models of such data capture spatial structure via the adjacency matrix A. Each element a_{ij} of it is either 1, if areas B_i and B_j are neighbours, and 0 if not. Point-referenced (geostatistical) data represent measurements of a spatially continuous phenomenon at a set of fixed locations. Number of cases recorded at a set of hospitals with known geographical position is an example of such data. Modelling of the underlying continuous process relies on pairwise distances between observed and unobserved locations. GPs in such models use continuous kernels. Point pattern data consist of precise locations of events. An example of a point pattern is a collection of GPS coordinates of households of newly occurred malaria cases. One way of modelling point pattern data is to cover the entire study area with a fine computational grid and view each grid cell as a location. Modelling of the continuous underlying process is typically done using continuous GP kernels. All of the three types of data can be modelled using the proposed novel approach.

2.2. Latent Gaussian models

Suppose we are given outcome data $y_1, \ldots, y_n$, corresponding to a set of observed disjoint areas $\{B_i\}$, $i = 1, \ldots, n$, covering the domain of interest $G = \cup_{i=1}^n B_i$. These areas may be the cells of a computational grid, pixels of a raster image, or they may correspond to existing administrative units. Outcome data $\{y_i\}_{i=1}^n$ can represent either counts aggregated over an area, such as number of disease cases, continuous bounded data, such as disease prevalence (i.e. a number between zero and one), or continuous unbounded data, such as land surface temperature in degrees Celsius (°C). Hence, in applied fields, it is common to

use generalized linear models to unify the modelling approach for all such types of outcome. Accounting for mixed effect model structure, their Bayesian hierarchical formulation can be expressed as follows:

$$\theta \sim p(\theta), \tag{2.1}$$

$$f|\theta \sim \mathrm{GP}(\mu, \Sigma(\theta)), \tag{2.2}$$

$$\eta = X\beta + f \tag{2.3}$$

and
$$y|\eta \sim p(u^{-1}(\eta), \theta). \tag{2.4}$$

Here, (2.1) describes the hyperparameters, θ, of the model, f in (2.2) denotes the latent Gaussian field defined by mean μ and covariance $\Sigma(\theta)$, X in (2.3) is the fixed effects design matrix (a set of covariates), β are the fixed effects and η is the linear predictor combining fixed and random effects. Equation (2.4) provides an observational model, where u is a link function characterizing the mean of the distribution (e.g. logit for binomial data, exponential for positive data). A common modelling assumption is that the observations y_i are conditionally independent $p(y|f, \theta) = \prod_{i=1}^{n} p(y_i|\eta(f_i), \theta)$, and the spatial structure is captured latently by function f. It is common to choose a GP prior over f, and as a consequence, finite realizations f_{GP} are jointly normally distributed with mean μ and covariance matrix Σ. Since f_{GP} always enters the model via the linear predictor, without loss of generality (affine transformation), we can consider f_{GP} to have zero mean ($\mu = 0$) and be distributed as $f_{\mathrm{GP}} \sim \mathcal{N}(0, \Sigma)$. Structure of the covariance matrix Σ depends on the spatial setting of the problem and the model which we chose for the random effect. Once the model for f_{GP} has been defined, the linear predictor can be computed as

$$u(E[y|f_{\mathrm{GP}}]) = X\beta + f_{\mathrm{GP}}, \tag{2.5}$$

and then linked to the observed data via the likelihood. Unless the random effect is chosen to be trivial (i.e. i.i.d. set of variables resulting from $\Sigma = I$), the random effect f_{GP} represents the computational challenge. Further, we describe some options for spatial random effect priors in the context of small-area estimation, and propose a method to substitute its calculation at the inference stage with another variable, leading to increased inference efficiency.

2.3. Models of areal data

A widely adopted group of approaches relies on the neighbourhood/adjacency structure and defines Σ based on the connectivity of the adjacency graph. Such methods leverage the tendency for adjacent areas to share similar characteristics. Conditional auto-regressive (CAR) and intrinsic conditional auto-regressive (ICAR) models were first proposed by Besag [20] and later extended to the Besag–York–Mollié (BYM) model [21]. The general form of the spatial prior for this group of models is

$$\phi \sim \mathcal{N}(0, Q^{-1}),$$

where the precision matrix Q defines which specific model is used. Under the CAR model, random effect $f_{\mathrm{CAR}} = \phi$ is defined by the prior $\phi \sim \mathcal{N}(0, \tau^{-1}R^{-})$. Here by $Q^{-} = \tau^{-1}R^{-}$, we denote generalized inverse [22] of the precision matrix Q. The precision matrix is calculated as $Q = \tau(D - \alpha A)$. Here, A is the adjacency matrix, and D is a diagonal matrix, with elements d_{ii} given by the total number of neighbours of area B_i. The parameter τ is the marginal precision, and the parameter α controls the amount of spatial dependence. For instance, $\alpha = 0$ implies complete independence and i.i.d. random effects. Condition $|\alpha| < 1$ ensures that the joint distribution of ϕ is proper [23]. It is not uncommon, however, to use $\alpha = 1$ leading to the degenerate precision matrix (hence, the Q^{-} notation) and the ICAR model. In practice, ICAR models are supplied with additional constraints,

such as $\sum_{i=1}^{n} \phi \approx 0$ to enable inference. The BYM model includes both an i.i.d. random effect component $\phi_1 \sim \mathcal{N}(0, \tau_1^{-1}I)$ to account for non-spatial heterogeneity and an ICAR component $\phi_2 \sim \mathcal{N}(0, Q_2^{-})$, $Q_2 = \tau_2 R_2$, for spatial auto-correlation. Hence, the total random effect can be calculated as $f_{\mathrm{BYM}} = \phi_1 + \phi_2$. Some reparameterizations of BYM have been recently proposed in the literature [24] to improve the interpretability of the inference. The advantage of the set of models presented above is that they take neighbourhood structure into account. However, neither the shapes (irregularity) nor the sizes of areas are captured.

Another natural way to model areal data, especially gridded surfaces with small-area sizes, is modelling covariance between areas as an approximately spatially continuous process, for example based on the pairwise distances between the centroids. Typical kernels for distance-based covariance matrices include squared exponential, exponential, periodic or Matérn.

2.4. Variational autoencoders

An autoencoder is a neural network that is trained by unsupervised learning, with the aim of dimensionality reduction and feature learning. It consists of an encoder network, $\mathcal{E}_\gamma(\cdot)$ with parameters γ, a decoder network $\mathcal{D}_\psi(\cdot)$ with parameters ψ and a bottleneck layer containing a latent vector z. The encoder maps input $x \in \mathcal{R}^p$ into the latent vector $z \in \mathcal{R}^d$, and the decoder network maps z back into the input space by creating a reconstruction of the original data $\hat{x} = \mathcal{D}_\psi(\mathcal{E}_\gamma(x))$. The network is trained by optimization to learn reconstructions $\hat{x}$ that are close to the original input x. An autoencoder can borrow any neural network architecture, including multilayer perceptrons, convolutional or recurrent layers. A VAE is a directed probabilistic graphical model whose posterior is approximated by a neural network, forming an autoencoder-like architecture. The goal of VAEs is to train a probabilistic model in the form of $p(x, z) = p(x|z)p(z)$, where $p(z) = \mathcal{N}(0, I)$ is a prior distribution over latent variables z and $p(x|z)$ is the likelihood function that generates data x given latent variables z. The output of the encoder network $\mathcal{E}_\gamma$ in VAEs is a pair of d-dimensional vectors $\mu_\gamma(x)$, $\sigma_\gamma^2(x)$, which can be used to construct the variational posterior for latent variable z. The decoder (generator) network $\mathcal{D}_\psi$ tries to reconstruct the input by producing $\hat{x}$. In particular, the model can be summarized into the following:

$$(\mu_\gamma(x), \sigma_\gamma^2(x)) = \mathcal{E}_\gamma(x), \tag{2.6}$$

$$z|x \sim \mathcal{N}(\mu_\gamma(x), \sigma_\gamma^2(x)I) \tag{2.7}$$

and
$$\hat{x}|z \sim \mathcal{D}_\psi(z), \quad z \sim \mathcal{N}(0, I). \tag{2.8}$$

Neural network parameters γ and ψ are estimated as maximizers of the evidence lower bound (ELBO)

$$\mathcal{L} = p(x|z, \gamma, \psi) - \mathrm{KL}(\mathcal{N}(\mu_\gamma(x), \sigma_\gamma^2(x)I)\|\mathcal{N}(0, I)),$$

or its extensions [25]. The first term in ELBO is the reconstruction loss, measured by likelihood quantifying how well $\hat{x}$ and x match. The second term is a KL divergence which ensures that z is as similar as possible to the prior distribution, a standard normal. It serves as a regularizer ensuring parsimony in the latent space and thus leads to uncorrelated variables in the latent space. New data can be generated by sampling from the latent space with the trained decoder network. Once the numerically optimal network parameters $\hat{\psi}$ have been obtained, new realizations can be generated in two steps. As the first step, we draw from the standard normal distribution $z \sim \mathcal{N}(0, I)$, and as the second step, apply the deterministic transformation $\mathcal{D}_{\hat{\psi}}(z)$.

2.5. The proposed method

Most commonly, VAEs have been used in the literature to learn and generate observed data. We propose using spatial priors
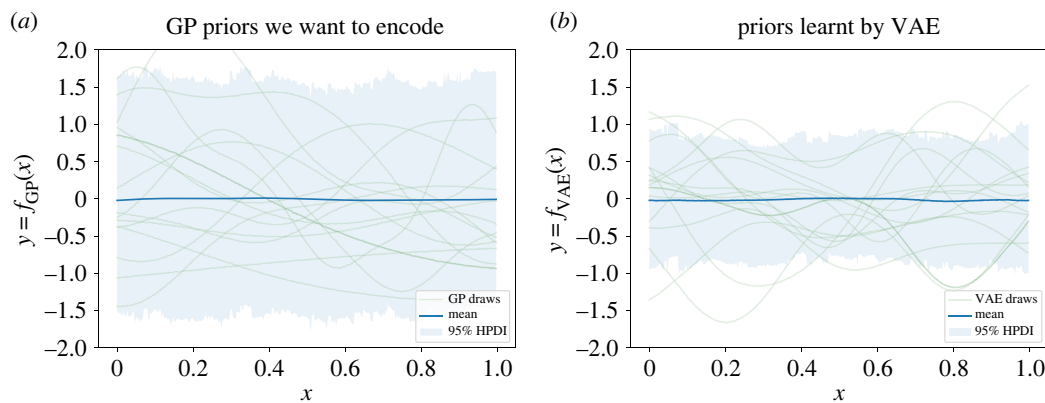
Figure 1. Learning one-dimensional GP priors on a regular grid with VAE: (*a*) prior samples from the original Gaussian process evaluations f_{GP}, (*b*) prior draws from f_{VAE} trained on f_{GP} draws. Mean and highest posterior density interval (HPDI) are calculated from 1000 samples.

$x = f_{GP}$, evaluated at the required spatial configuration (a grid or neighbourhood structure), as training examples instead of observed data. The trained VAE then enables us to compute $\hat{x} = f_{VAE}$. Remarkably, unlike dealing with observed data, this approach does not have issues with the quality nor the quantity of the training examples. The amount of training data is unlimited since we can draw as many GP realizations for training as required. Similarly, there is no issue of data quality as we can create exact GP draws, free from noise, characteristic for any real-life observations.

To perform MCMC inference in a spatial model, we replace evaluation of the GP prior f_{GP} in the linear predictor (2.5) with the learnt prior f_{VAE} at the inference stage:

$$u(E[y|f_{VAE}]) = X\beta + f_{VAE}.$$

Drawing from the standard normal distribution $z \sim \mathcal{N}(0, I)$ with uncorrelated entries z_i leads to a much higher efficiency when compared with the highly correlated multivariate normals $\mathcal{N}(0, \Sigma)$ with a dense covariance matrix Σ. As a consequence, computation time also decreases, especially for models where $d < p$.

3. Results

3.1. One-dimensional Gaussian process on regular and irregular grids

In this first example, we use VAE to perform inference on continuous data $\{y_i\}_{i=1}^n$ over a regular one-dimensional grid. The grid consists of $n = 400$ points in the $(0, 1)$ interval. Training prior samples are drawn as evaluations of a GP with zero mean and squared exponential kernel $k(h) = \sigma^2 e^{-h^2/l^2}$. This model is useful for small-area estimation when the covariance matrix is Euclidean distance-based. To allow for hyperparameter learning, we impose hierarchical structure on the model by using hyperpriors on variance $\sigma^2 \sim \text{LogNormal}(0, 0.1)$ and lengthscale $l \sim \text{InverseGamma}(4, 1)$. This hierarchical structure allows the VAE to be trained on a range of values of hyperparameters. The average lengthscale, according to these priors, is around 0.3. Since this is much larger than the distance between two neighbouring points on the grid (0.0025), there is enough redundancy in the data to expect a lower-dimensional embedding. Realizations f_{GP} are presented in figure 1*a*. We trained a VAE with two hidden layers of dimensions 35 and 30, respectively, and the bottleneck layer of dimension 10. As an activation function we used the rectified linear unit [26–28] for all nodes in the hidden layers. The priors learnt by the VAE are presented in figure 1*b*. Figure 1 shows that

the typical shape of the prior as well as the mean have been learnt well. Amount of uncertainty (second moment) displayed by the VAE priors is lower. This is an expected phenomenon as vanilla VAEs are known to produce blurred and over-smoothed data due to compression; on the other hand, they perform well on denoising tasks [29], i.e. reconstructing underlying truth given corrupt data. Empirical covariance matrices of the original and trained VAE priors show similar patterns (electronic supplementary material, figure S12). To perform inference, we generate one GP realization, use it as the ground truth, and simulate observed data by adding i.i.d. noise. We allow the number of observed data points to vary as 0.5% (2 data points), 1% (4 data points) and 1.5% (6 data points) of the total number of the grid points and recover the true function. The model used for inference is $y \sim \mathcal{N}(f_{VAE}, s^2)$, where the amount of noise is given the half-normal prior $s \sim \mathcal{N}^+(0.1)$. Inference results are presented in figure 2. The higher the number of points, the closer is the estimated mean to the ground truth curve. Areas without any data available in their proximity show higher uncertainty than areas informed by closely located data points. Effective sample size (ESS) is an important measure of the efficiency of MCMC sampling [30]. For example, we have run inference with 1000 warm-up and 1000 posterior MCMC samples for different number of data points. Average ESS for the posterior of the function evaluated at the observed points increased together with the number of points, while inference time remained constant. The original GP model displayed the reverse trend: average ESS remained constant, while computation time increased.

To verify that our approach is also applicable to irregular grids, we modified the simulation study by placing the grid points irregularly. We use the same architecture to train this model, i.e. 35, 30 nodes in each hidden layer and 10 latent variables in the bottleneck layer. Original GP priors evaluated on an irregular grid, the learnt VAE priors and inference results are presented in electronic supplementary material, figures S13 and S14.

3.2. Two-dimensional Gaussian process

A similar set of experiments as described above were performed for two-dimensional GP priors. A regular grid was defined over a unit square with 25 segments along each of the two coordinates, resulting in $n = 625$ grid cells. Inference results produced using a VAE trained on evaluations of two-dimensional GP priors are presented in figure 3. As the number of observed
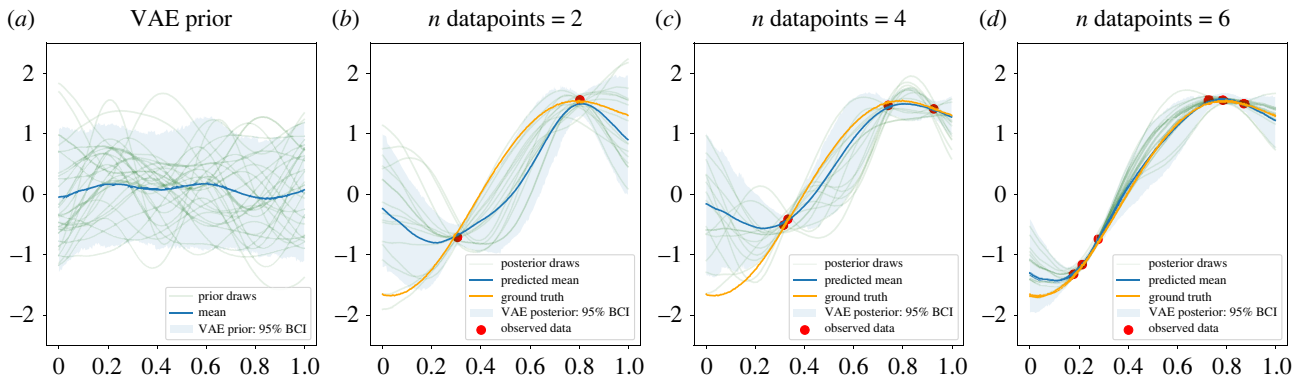
Figure 2. We perform MCMC inference on data generated from a noise-free GP and added i.i.d. noise by using the trained VAE priors f_{VAE}. (a) VAE prior. (b–d) Posterior mean of our model in green with the 95% credible intervals shown in blue. Quality of the estimation improves with the growing number of data points.
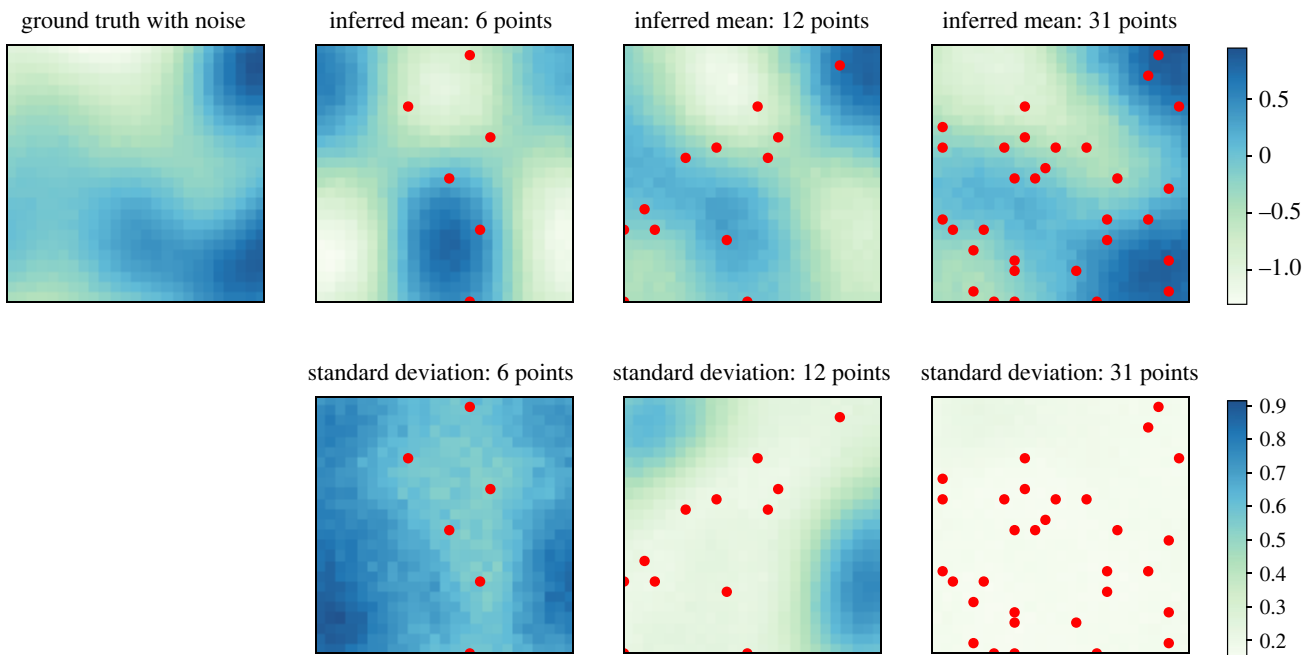

Figure 3. Inference results on a two-dimensional grid using the VAE-based priors. For comparison, inference has been performed using 1%, 2% and 5% of the total number of points (6, 12 and 31 points, respectively). Uncertainty in the surface estimates decreases as the number of observed points increases.

points increases, quality of the mean prediction improves and uncertainty in the surface estimates decreases.

3.3. Synthetic conditional auto-regressive example

We partition a rectangle by subdividing it into 10 rows and 15 columns. Unlike the two examples above, where we work with a continuous kernel, here we generate data using the CAR model $\phi \sim N(0, Q^{-1})$, $Q = \tau(D - \alpha A)$, reliant on the adjacency matrix A. Parameter α is being assigned the Uniform(0.4, 1) prior. This prior is chosen to study the question whether VAE can be trained well on CAR samples, displaying spatial correlation which corresponds to the values of α away from zero.

In this example, we demonstrate a technique to train a VAE by allowing it to preserve explicit estimation of the marginal precision τ. For this VAE, the training is performed on the standardized samples $\bar{\phi}$ from the random effect, meaning that the prior is drawn from the distribution

$$\bar{\phi} \sim N(0, \bar{Q}^{-1}),$$
$$\bar{Q} = D - \alpha A. \tag{3.1}$$

Since $\phi = (1/\sqrt{\tau})\bar{\phi} \sim N(0, Q^{-1})$, the trained VAE generator $\bar{f}_{VAE-CAR}$ also needs to be adjusted in the model at the inference stage by the magnitude of the marginal precision τ: $f_{VAE-CAR} = (1/\sqrt{\tau})\bar{f}_{VAE-CAR}$.

To test the performance of the trained VAE, ground truth data are generated by fixing the value of α at 0.7, and adding normal i.i.d. noise term at each area with variance of 0.5. These data are then used to fit the original CAR model, as well as the model using a VAE trained on CAR prior samples. Both models use the prior Uniform(0.01, 1) for the variance of the noise term. For both models, we run MCMC with 1000 warm-up iterations and 2000 iterations to sample from posterior. The average ESS of the spatial random effect in the VAE–CAR model is 4650 which took 9 s to compute; the achieved mean squared error (MSE) is 0.135. The average ESS in the original CAR model is 3128 which took 73 s; the achieved MSE is 0.118. Results of the experiment are presented in figure 4.

3.4. Scottish lip cancer dataset

The Scottish lip cancer dataset, originally presented by Kemp et al. [31], has become a benchmark dataset for areal models. It
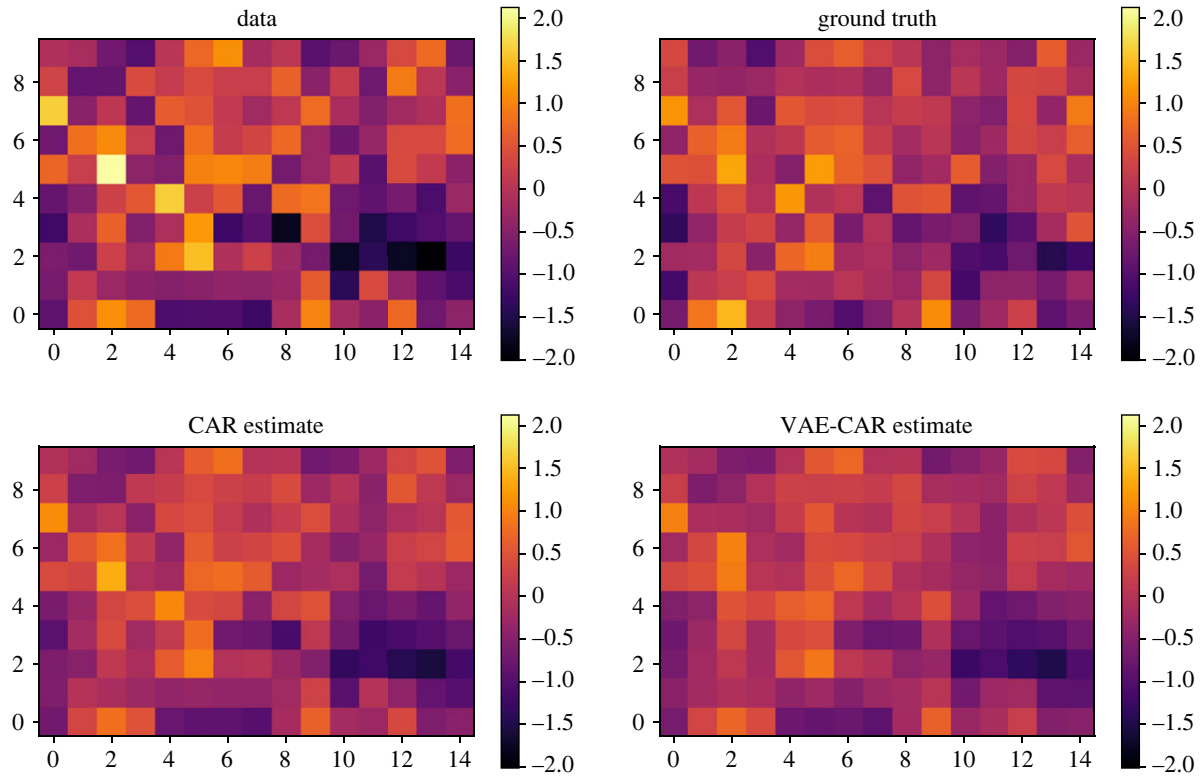
Figure 4. Inference results on a synthetic CAR example: data are generated by adding a noise term to the ground truth. We perform MCMC inference on the data by using both the original CAR and the VAE–CAR models.

has been used to demonstrate performance and implementations of CAR, ICAR, BYM and its variations [32,33]. The dataset consists of the observed and expected numbers of cases (y and E, respectively) at 56 counties in Scotland, as well as a covariate measuring the proportion of the population engaged in agriculture, fishing or forestry (*aff*). The covariate is related to exposure to sunlight, which is a risk factor for lip cancer. We model the count data y as following a Poisson distribution with the log-Normal rate λ distributed according to the BYM model:

$$y \sim \text{Poisson}(\lambda),$$
$$\log(\lambda) = \log(E) + b_0 + b_1 \, \text{aff} + \phi_1 + \phi_2,$$
$$\phi_1 \sim \mathcal{N}(0, \tau_1 I)$$

and $\qquad \phi_2 \sim \mathcal{N}(0, Q_2^-), \quad Q_2 = \tau_2(D - A).$

The VAE is trained on the spatial random effect BYM priors $f_{\text{BYM}} = \phi_1 + \phi_2$ to obtain the $f_{\text{VAE-BYM}}$ representation. We can use any parameterization of BYM, as we only rely on its generative properties (and this is one of the advantages of our approach). We note that a model with i.i.d. random effect ϕ_1 already produces a relatively good fit (see electronic supplementary material, figure S15a). It is only the remaining discrepancies that the spatially structured random effect ϕ_2 needs to explain. We account for this observation in our priors for τ_1 and τ_2, as well as the dimension of the latent variable z: as there is only a moderate spatial signal, there is no redundancy in the data for spatial effect estimation. To be able to provide good quality VAE–BYM priors, we opt to not compress the spatial prior and choose the dimension of z to be equal to the number of counties. We train a network with one hidden layer with 56 hidden nodes and use the exponential linear unit [34], or *elu*, activation function. For optimization, we use the variational Rényi bound that extends traditional variational inference to Rényi α-divergences as proposed by Li & Turner [25]. By using $\alpha = 0$, we opt for an
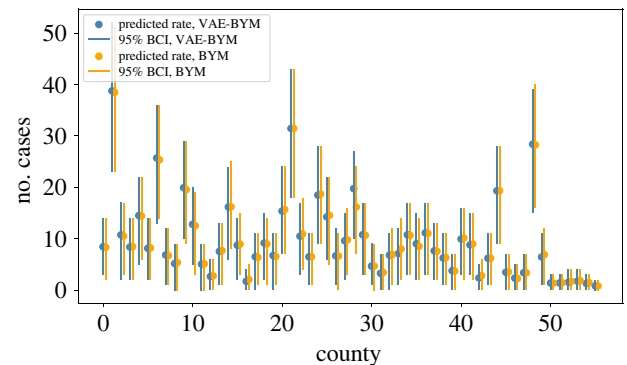

Figure 5. Scottish lip cancer dataset: posterior predictive distributions of the rates λ_{BYM} and $\lambda_{\text{VAE-BYM}}$ produced by models with f_{BYM} and $f_{\text{VAE-BYM}}$ random effects, respectively. The distributions are represented by the point estimates (mean) and uncertainty intervals (95% BCI) for each county. There is a very good agreement between the two models.

importance weighted autoencoder [35]. We performed two assessments to evaluate whether the VAE–BYM produces similar inference results to the original BYM model. First, we used both models for inference on the entire dataset to compare results for mapping (the most typical task in epidemiology and policy informing work), and second, we performed fivefold cross-validation. Posterior predictive distributions of the rates λ_{BYM}, $\lambda_{\text{VAE-BYM}}$ obtained by the models, where f_{BYM} and $f_{\text{VAE-BYM}}$ have been used to capture the spatial random effect, are very close to each other: figure 5 displays the two distributions, where each of them is represented by its point estimate (mean) and 95% Bayesian credible interval (BCI). Uncertainty intervals produced by the two models are remarkably close to each other for most of the counties. Figure 6 demonstrates very good agreement in the point estimates. Figure 7 presents the obtained maps: models with f_{BYM} and
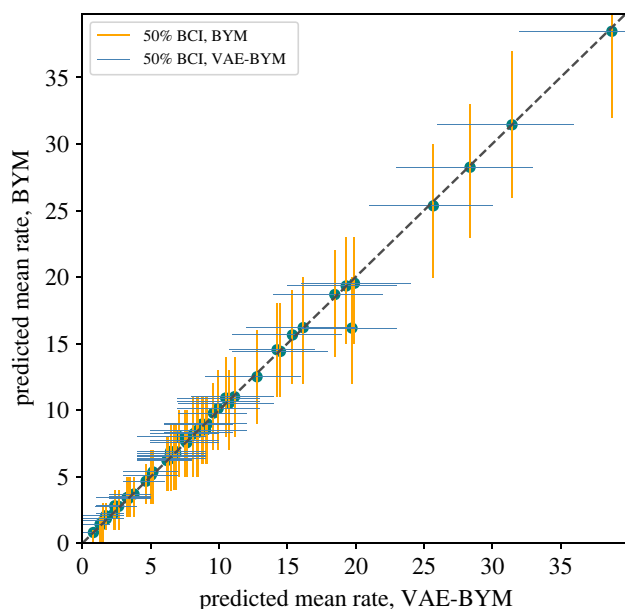
Figure 6. Scottish lip cancer dataset: point estimates (means) and uncertainty intervals (50% BCIs) of the rates λ_{BYM} and $\lambda_{VAE-BYM}$ produced by models with f_{BYM} and $f_{VAE-BYM}$ random effects, respectively. There is a very good agreement between the two models.

$f_{VAE-BYM}$ produce very close spatial patterns. The average ESS of the spatial effects in the BYM model is approximately 150, and in the VAE–BYM model it is approximately 1030. MCMC elapsed time shows the same trend: 402 s and 12 s for the BYM and VAE–BYM, respectively. To perform cross-validation, we created a fivefold split of the data. To measure performance, we have used MSE between the continuous predicted rate λ and the observed count y. The mean MSE of the BYM model across five runs was 426, with standard deviation of 131, and the mean MSE of the VAE–BYM model was 414, with standard deviation of 171. Average ESS of the VAE–BYM random effect was approximately 3850, and average ESS of the BYM random effect was approximately 630. Inference times were 3 s on average (0.2 s standard deviation) for the VAE–BYM runs, and 33 s on average (3 s standard deviation) for the BYM runs. These experiments confirm the consistency of our observations: even when the dimension of the latent variable z is the same as the number of the counties, there is a benefit to using VAE–BYM over BYM—it achieves comparable performance while displaying much higher ESS and shorter inference times.

3.5. An alternative approach to training a variational autoencoder for non-Gaussian likelihoods

Our previously described approach, where we calculate the linear predictor and then fit model to the data using a link function, follows the long standing tradition in statistics driven by the interest to measure associations between predictors and an outcome. In the previous two examples, we have trained the VAE directly on the GP draws (f_{GP}) and used the Gaussian distribution to compute the reconstruction loss $p(x|\hat{x})$. When the outcome data are not Gaussian or not even continuous, there is an alternative way of training a VAE. Let us consider, for instance, count data which we would like to model using the Poisson distribution. Instead of using f_{GP} draws as training data of the VAE, we can use directly the simulated counts y arising from the Poisson distribution with rate $\lambda = \exp(f_{GP})$. The encoder now maps

these counts into the latent variable z, as before, and the decoder is enhanced with a probabilistic step: it maps z into a positive rate $\hat{\lambda}$, and calculates the reconstruction loss as $p(y|\hat{\lambda}) : y \sim \text{Pois}(\hat{\lambda})$. Results of such a model are shown in figure 8. We have generated training data y for the VAE over a regular grid with $n = 100$ points according to the distribution $y \sim \text{Pois}(\exp(f_{GP}))$. The hyperparameters of the GP were identical to our one-dimensional GP example presented above. VAE architecture included two hidden layers with *elu* activation function. An exponential transform was applied to the final layer of the decoder to calculate the predicted rate $\hat{\lambda}$.

3.6. HIV prevalence in Zimbabwe

We consider household survey data from the 2015 to 2016 Population-based HIV Impact Assessment Survey in Zimbabwe [36]. The observed positive cases y_i among all observed cases n_i in each area B_i are modelled according to the Binomial distribution:

$$y \sim \text{Binomial}(n, \theta),$$

$$\text{logit}^{-1}(\theta) = b_0 + \phi,$$

$$\phi \sim \mathcal{N}(0, Q^{-1})$$

and

$$Q = \tau(D - \alpha A).$$

The parameter θ_i is the 'probability of success' of the binomial distribution, and serves as the estimate of HIV prevalence. The linear predictor $\text{logit}^{-1}(\theta_i) = b_0 + \phi_i$, $i = 1, \ldots, N$ consists of the intercept b_0, common for each area, and area-specific spatial random effect ϕ_i. The VAE is trained on the spatial random effect CAR priors to obtain the VAE–CAR representation. We train VAE on the standardized draws of the spatial random effect as presented in formulae (3.1). The total number of observed areas is 63, and we trained a VAE with a dimension of 50 in the latent space to achieve reduction in the dimensionality of the latent variable. Figure 9 presents the obtained maps: models with f_{CAR} and $f_{VAE-CAR}$ produce very close spatial patterns. We used 2000 MCMC iterations to perform inference using both the original CAR, as well as the VAE–CAR models. The average ESS of the spatial effects in the CAR model is approximately 120 with a computation time of 13 s. In the VAE–CAR model average ESS is higher, at approximately 2600, and took only 4 s to compute. ESS achieved by the VAE–CAR model is about 20 times higher than the one achieved by the CAR model, and also higher than the actual number of posterior samples, which indicates extremely efficient sampling. Traceplots obtained by both models are presented in figure 10 and demonstrate (together with electronic supplementary material, figure S17) that posterior samples produced by the VAE–CAR model display very little auto-correlation, which allows to achieve high ESS and fast inference. Posterior samples produced by the original CAR model, on the contrary, show high auto-correlation (electronic supplementary material, figure S18), leading to low ESS and longer computation time.

3.7. COVID-19 incidence in the UK

We consider the projected number of COVID-19 infections in the UK at Lower Tier Local Authority (LTLA) level [37]. These estimates are available from a public website[1] and are based on the Scott *et al.* [38] package. The model behind the package [39] relies on a self-renewal equation and does
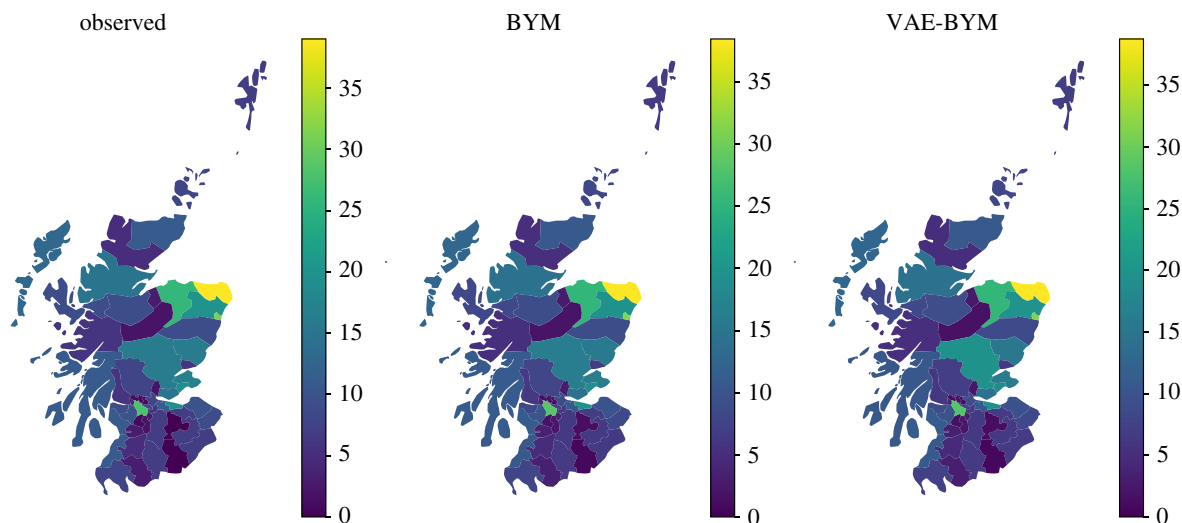
Figure 7. We perform MCMC inference on the observed number of lip cancer cases in each of the 56 counties of Scotland. The leftmost plot displays the observed number of cases, the middle plot shows the mean of the posterior predictive distribution using the BYM random effect f_{BYM}, the rightmost plot shows the mean of the posterior predictive distribution obtained from the model with the $f_{\text{VAE-BYM}}$ random effect. Models using BYM and VAE–BYM display the same pattern of the disease spatial distribution.
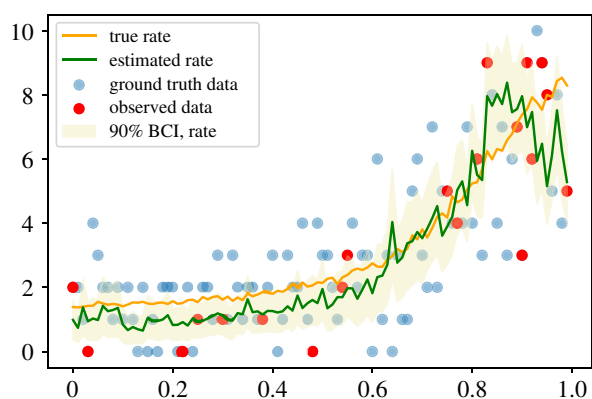
Figure 8. Inference results obtained via a VAE trained directly on the count data y generated from the Poisson distribution $\text{Pois}(\lambda)$, $\lambda = \exp(f_{\text{GP}})$. The estimated rate is in good agreement with the ground truth.

not take spatial correlations into account. Here, we demonstrate how spatial modelling using the proposed approach can be performed on the incidence data. Number of infections y_i in each LTLA during the period between 21 January and 5 February 2022 is modelled via the Poisson distribution according to the model

$$y \sim \text{Poisson}(\lambda),$$
$$\log(\lambda) = b_0 + \phi,$$
$$\phi = \frac{1}{\sqrt{\tau}} \bar{\phi},$$
$$\bar{\phi} \sim \mathcal{N}(0, \bar{Q}^{-1})$$
and
$$\bar{Q} = D - \alpha A.$$

As above, we have chosen the representation of the spatial CAR random effect in its standardized form $\bar{\phi}$ in order to allow for explicit inference of the marginal precision τ when using VAE-based inference. To simplify modelling, we have removed all singletons (islands without any neighbouring areas within them) from the shapefile. We trained VAE on the standardized draws of the spatial random effect. The

total number of modelled LTLAs is 372, and we chose the dimension of the latent space to be 300. Using this example, we demonstrate a technique, where several VAEs using different priors for the spatial hyperparameter α can be pre-trained, and then used for model selection—a step enabled by fast inference times when using VAE-based models. We have trained five VAEs using hyper-priors $\alpha \sim U(0, 0.2)$, $U(0.2, 0.4)$, $U(0.4, 0.6)$, $U(0.6, 0.8)$, $U(0.8, 0.99)$, correspondingly. Each of the resulting VAE priors were used to fit a model. Model selection was performed based on the widely applicable information criterion (WAIC) [40] using the *arviz* package [41]. The best model was the one trained with $\alpha \sim U(0.8, 0.99)$. To obtain smooth maps, we used 2000 MCMC iterations and performed inference using both the original CAR, as well as the VAE–CAR models with priors $\alpha \sim \text{Uniform}(0.8, 0.99)$ and $\tau \sim \text{Gamma}(6, 2)$. Resulting maps are presented in figure 11: models with f_{CAR} and $f_{\text{VAE-CAR}}$ produce similar spatial patterns. Characteristics of the two model fits are presented in table 1: there is no significant difference between the absolute errors (p-value of the paired t-test is 0.05), while VAE–CAR has achieved much higher ESS at much shorter computation times.

4. Discussion and future work

In this paper, we have proposed a novel application of VAEs to learn spatial GP priors and enable small-area estimation. Such an approach leverages the power of deep learning to fuel inference for well-established statistical models. Uncorrelated latent parameters of the VAE make consequent Bayesian inference with MCMC highly efficient at successfully exploring the posterior distribution. An advantage of the proposed approach, when compared with traditional VAE applications, is that there are no limitations in neither quality nor quantity of the training data as any number of training samples can be generated by drawing from the noise-free spatial priors. In addition, there is no need to retain the whole training dataset as every batch can be generated on the fly. Our method is beneficial even if the latent space dimension is the same as
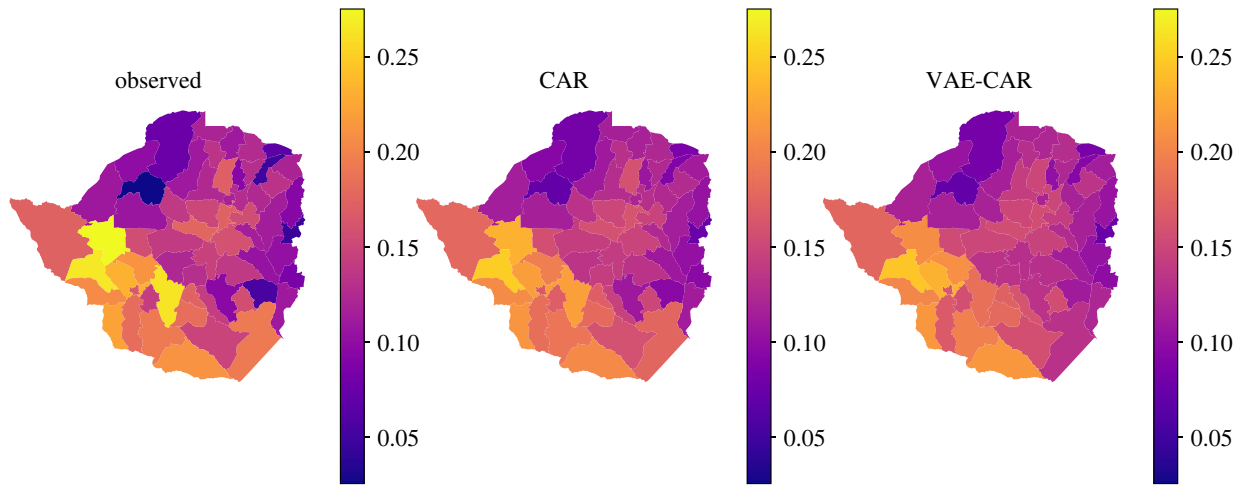
Figure 9. We perform MCMC inference on the observed number of HIV-positive individuals among observed individuals in Zimbabwe. The leftmost plot displays raw data, the middle plot shows the mean of the posterior predictive distribution using the CAR spatial random effect f_{CAR}, the rightmost plot shows the mean of the posterior predictive distribution obtained from the model with the $f_{VAE-CAR}$ spatial random effect.
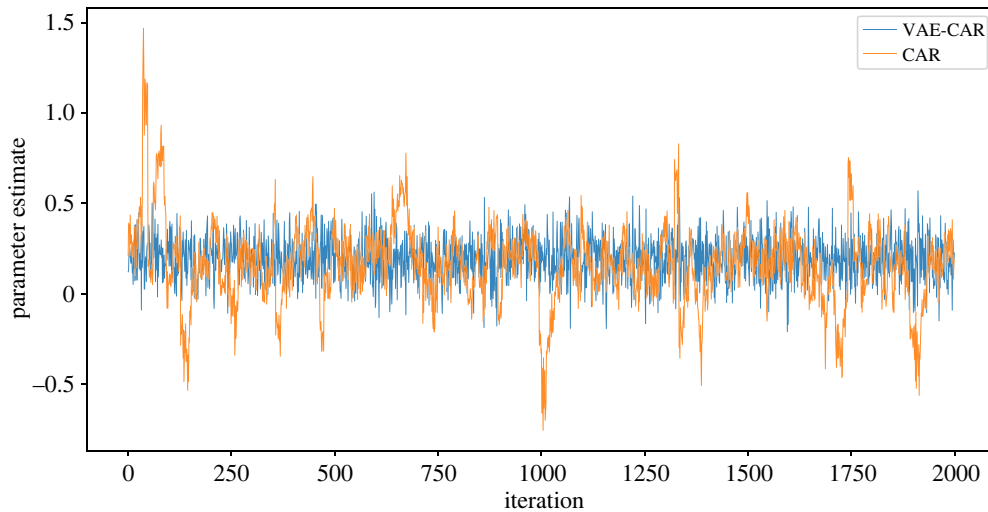

Figure 10. Traceplots of the spatial random effect for one area obtained during MCMC inference on HIV prevalence data from Zimbabwe. Posterior samples produced by the VAE–CAR model display very little auto-correlation, which allows to achieve high ESS and fast inference. Posterior samples produced by the original CAR model, on the contrary, show high auto-correlation, leading to low ESS and longer computation time.

the number of data points, unlike some other approximation methods, which rely on information redundancy in the data. As the HIV prevalence in Zimbabwe example shows, by reducing the auto-correlation between posterior MCMC samples, we can obtain drastic gains in both ESS and computation speed.

The limitations of the proposed approach are as follows. Firstly, MCMC inference is restricted to the spatial structure used for training. For instance, if a fixed spatial grid or a neighbourhood structure was used to train the VAE, prediction for off-grid points would not be possible. Secondly, using the MSE loss for VAE training, we do not expect the VAE to work well for values of hyperparameters outside of the typical ranges of hyperpriors which the VAE was trained on: if a VAE was trained on GP realizations with short lengthscales, it is unreasonable to expect good results for long lengthscales. VAE training losses which capture distributional properties of the training samples have the potential to resolve it and constitute future work. The over-smoothing property of vanilla VAEs can be accounted for at

the inference stage via priors allowing for wider uncertainty ranges. There is also the upfront cost of training a VAE, including the choice of architecture. Finally, recovering interpretable hyperparameters of the spatial priors might be of scientific interest and is more challenging in the VAE approach rather than the traditional framework. In some cases, this difficulty can be alleviated. For instance, in the HIV prevalence example, we have demonstrated how marginal precision or variance can be isolated from the VAE training to retain direct inference of this parameter.

Statistical models of areal data described above assign the same characteristics to each location within every area B_i. This assumption is unrealistic for many settings as heterogeneity might be present within each B_i, as long as the size of the area is non-negligible. Areal data can be viewed as an aggregation of point data and a series of approaches began to emerge recently to address this issue [42,43]. Hence, it is reasonable to use a data augmentation step to sample from the posterior distribution of the exact locations, and then to aggregate results. Future directions of research include an
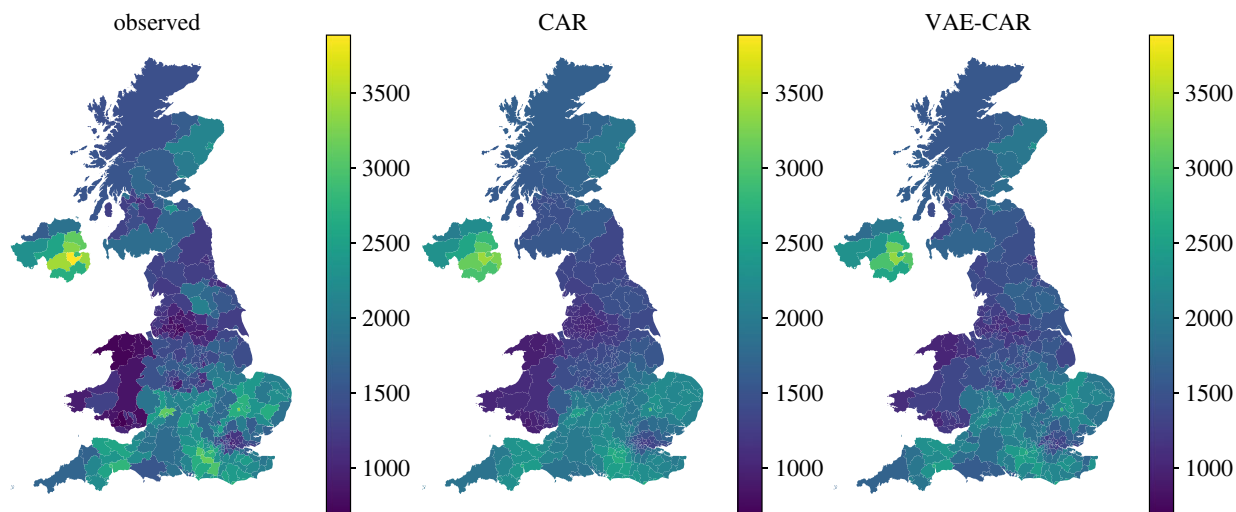
Figure 11. We perform MCMC inference on the projected COVID-19 incidence in the UK at the LTLA level. The leftmost plot displays raw data, the middle plot shows the mean of the posterior predictive distribution using the CAR spatial random effect f_{CAR}, the rightmost plot shows the mean of the posterior predictive distribution obtained from the model with the $f_{VAE-CAR}$ spatial random effect.

Table 1. Results of MCMC inference with CAR and VAE–CAR models: after 2000 iterations, VAE–CAR models achieve much higher average ESS at much shorter computation time, while displaying similar goodness of fit to the original model. *Mean and standard error.

model	absolute error*	ESS of the spatial random effect*	MCMC elapsed time (s)
CAR	1.53 (0.06)	317 (6)	277
VAE–CAR	1.63 (0.06)	3188 (24)	8

approach where models presented above, reliant on the adjacency structure, are substituted with continuous kernel approaches and a VAE trained using them.

The application of VAEs to small-area estimation has potential for far-reaching societal impact. If a set of different GP priors, such as CAR, ICAR, BYM and others, are used to pretrain several VAEs over the same spatial configuration, the resulting decoders can then be applied via the proposed inference scheme to rapidly solve real-life problems. Once the VAE training has been performed, users will only need access to the decoder parameters, and otherwise perform inference as usual—the expensive step of training a VAE would not be required at a practitioner's end. In case of an epidemic emergency, for instance, this would enable faster turnaround times in informing policy by estimating crucial quantities such as disease incidence or prevalence.

References

1. Rao JN, Molina I. 2015 *Small area estimation*. Hoboken, NJ: John Wiley & Sons.

2. Clements AC, Lwambo NJ, Blair L, Nyandindi U, Kaatano G, Kinung'hi S, Webster JP, Fenwick A, Brooker S. 2006 Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Trop. Med. Int. Health* **11**, 490–503. (doi:10.1111/j.1365-3156.2006.01594.x)

3. Williams CK, Rasmussen CE. 2006 *Gaussian processes for machine learning*, vol. 2. Cambridge, MA: MIT Press.

4. Stephenson WT, Ghosh S, Nguyen TD, Yurochkin M, Deshpande SK, Broderick T. 2021 Measuring the sensitivity of gaussian processes to kernel choice. (https://arxiv.org/abs/2106.06510)

5. Rue H, Martino S, Chopin N. 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B (Stat. Methodol.)* **71**, 319–392. (doi:10.1111/j.1467-9868.2008.00700.x)

6. Hoffman MD, Blei DM, Wang C, Paisley J. 2013 Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347.

7. Minka TP. 2013 Expectation propagation for approximate Bayesian inference. (https://arxiv.org/abs/1301.2294)

8. Titsias M. 2009 Variational learning of inducing variables in sparse gaussian processes. In *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009*, pp. 567–574.

9. Yao Y, Vehtari A, Simpson D, Gelman A. 2018 Yes, but did it work?: evaluating variational inference. In *Proc. 35th Int. Conf. on Machine Learning, Stockholm, Sweden, 10–15 July 2018*, pp. 5581–5590.

10. Martins TG, Simpson D, Lindgren F, Rue H. 2013 Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* **67**, 68–83. (doi:10.1016/j.csda.2013.04.014)

11. Carpenter B et al. 2017 Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32. (doi:10.18637/jss.v076.i01)

12. Phan D, Pradhan N, Jankowiak M. 2019 Composable effects for flexible and accelerated probabilistic programming in NumPyro. (https://arxiv.org/abs/1912.11554)

13. Salvatier J, Wiecki TV, Fonnesbeck C. 2016 Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2**, e55. (doi:10.7717/peerj-cs.55)

14. Ge H, Xu K, Ghahramani Z. 2018 Turing: a language for flexible probabilistic inference. In *Proc. 21st Int. Conf. on Artificial Intelligence and Statistics, Lanzarote, Spain, 9–11 April 2018*, pp. 1682–1690.

15. Fortuin V, Baranchuk D, Rätsch G, Mandt S. 2020 GP-VAE: deep probabilistic time series imputation. In *Proc. 23rd Int. Conf. on Artificial Intelligence and Statistics, 26–28 August 2020*, pp. 1651–1661.

16. Mishra S, Flaxman S, Berah T, Pakkanen M, Zhu H, Bhatt S. 2020 πVAE: encoding stochastic process priors with variational autoencoders. (https://arxiv.org/abs/2002.06873)

17. Hinton GE, Salakhutdinov RR. 2006 Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507. (doi:10.1126/science.1127647)

18. Kingma DP, Welling M. 2013 Auto-encoding variational Bayes. (https://arxiv.org/abs/1312.6114)

19. Liu Q, Allamanis M, Brockschmidt M, Gaunt AL. 2018 Constrained graph variational autoencoders for molecule design. (https://arxiv.org/abs/1805.09076)

20. Besag J. 1974 Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B (Methodol.)* **36**, 192–225.

21. Besag J, York J, Mollié A. 1991 Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20. (doi:10.1007/BF00116466)

22. James M. 1978 The generalised inverse. *Math. Gazette* **62**, 109–114. (doi:10.1017/S0025557200086460)

23. Gelfand AE, Vounatsou P. 2003 Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11–15. (doi:10.1093/biostatistics/4.1.11)

24. Riebler A, Sørbye SH, Simpson D, Rue H. 2016 An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25**, 1145–1165. (doi:10.1177/0962280216660421)

25. Li Y, Turner RE. 2016 Rényi divergence variational inference. (https://arxiv.org/abs/1602.02311)

26. Glorot X, Bordes A, Bengio Y. 2011 Deep sparse rectifier neural networks. In *Proc. 14th Int. Conf. on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011*, pp. 315–323.

27. Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y. 2009 What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th Int. Conf. on Computer Vision*, pp. 2146–2153. (doi:10.1109/ICCV.2009.5459469)

28. Nair V, Hinton GE. 2010 Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th Int. Conf. on Machine Learning, Haifa, Israel, 21–24 June 2010*, pp. 807–814.

29. Kovenko V, Bogach I. 2020 A comprehensive study of autoencoders, applications related to images. In *Int. Conf. on Information Technology and Interactions, Workshops Proc., Kyiv, Ukraine, 2–3 December 2020*, pp. 43–54.

30. Martino L, Elvira V, Louzada F. 2017 Effective sample size for importance sampling based on discrepancy measures. *Signal Process.* **131**, 386–401. (doi:10.1016/j.sigpro.2016.08.025)

31. Kemp I, Boyle P, Muir C, Cameron L. 1985 *Atlas of cancer in Scotland, 1975–1980: incidence and epidemiological perspective*. Lyon, France: International Agency for Research on Cancer.

32. Duncan EW, White NM, Mengersen K. 2017 Spatial smoothing in Bayesian models: a comparison of weights matrix specifications and their impact on inference. *Int. J. Health Geogr.* **16**, 1–16. (doi:10.1186/s12942-017-0120-x)

33. Morris M, Wheeler-Martin K, Simpson D, Mooney SJ, Gelman A, DiMaggio C. 2019 Bayesian hierarchical spatial models: implementing the Besag York Mollié model in stan. *Spatial Spatio-temporal Epidemiol.* **31**, 100301. (doi:10.1016/j.sste.2019.100301)

34. Clevert D-A, Unterthiner T, Hochreiter S. 2015 Fast and accurate deep network learning by exponential linear units (elus). (https://arxiv.org/abs/1511.07289)

35. Burda Y, Grosse R, Salakhutdinov R. 2015 Importance weighted autoencoders. (https://arxiv.org/abs/1509.00519)

36. Sachathep K et al. 2021 Population-based HIV impact assessments survey methods, response, and quality in Zimbabwe, Malawi, and Zambia. *J. Acquir. Immune Defic. Syndr.* **87**, S6–S16. (doi:10.1097/QAI.0000000000002710)

37. Mishra S, Scott J, Zhu H, Ferguson NM, Bhatt S, Flaxman S, Gandy A. 2020 A COVID-19 model for local authorities of the United Kingdom. *medRxiv*. (doi:10.1101/2020.11.24.20236661)

38. Scott JA, Gandy A, Mishra S, Unwin J, Flaxman S, Bhatt S. 2020 Epidemia: modeling of epidemics using hierarchical Bayesian models. R package version 1.0.0.

39. Flaxman S et al. 2020 Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261. (doi:10.1038/s41586-020-2405-7)

40. Watanabe S, Opper M. 2010 Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594.

41. Kumar R, Carroll C, Hartikainen A, Martin O. 2019 Arviz a unified library for exploratory analysis of Bayesian models in python. *J. Open Source Softw.* **4**, 1143. (doi:10.21105/joss.01143)

42. Arambepola R, Lucas TC, Nandi AK, Gething PW, Cameron E. 2022 A simulation study of disaggregation regression for spatial disease mapping. *Stat. Med.* **41**, 1–16. (doi:10.1002/sim.9220)

43. Johnson O, Diggle P, Giorgi E. 2019 A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. *Stat. Med.* **38**, 4871–4887. (doi:10.1002/sim.8339)

44. Semenova E, Xu Y, Howes A, Rashid T, Mishra S, Flaxman S. 2022 PriorVAE: encoding spatial priors with variational autoencoders for small-area estimation. Figshare. (https://doi.org/10.6084/m9.figshare.c.6011431)