# Encoding spatial priors with VAEs for geospatial modelling.

Elizaveta Semenova,
University of Oxford

5th Workshop on Tractable Probabilistic Modeling
UAI 2022

## The map that changed how we fight outbreaks

- In 1854 London experienced a cholera outbreak.

# The map that changed how we fight outbreaks

- In 1854 London experienced a cholera outbreak.
- People believed, it was transmitted by 'miasma' (evil mist).

CHOLERA "TRAMPLES THE VICTOR & THE VANQUISH'D BOTH."

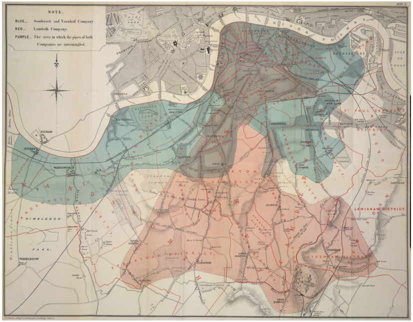
# The map that changed how we fight outbreaks

- John Snow has mapped cholera cases.

Credit: ESRI, "John Snow's cholera map"

# The map that changed how we fight outbreaks

- He also mapped different water companies' service areas.




Credit: British Library

# The map that changed how we fight outbreaks

- This pump has started a new field - spatial epidemiology.

# Outline

# Latent Gaussian models and GLMMs

$y = (y_1, ..., y_n)$ - outcome data over a set of $n$ locations

# Latent Gaussian models and GLMMs

$y = (y_1, ..., y_n)$ - outcome data over a set of $n$ locations

$y \sim p(y|g^{-1}(\eta), \theta)$ - observational model (likelihood)

# Latent Gaussian models and GLMMs

$y = (y_1, ..., y_n)$ - outcome data over a set of $n$ locations

$y \sim p(y|g^{-1}(\eta), \theta)$ - observational model (likelihood)

$\eta = X\beta + f$ - linear predictor: fixed and random effects

# Latent Gaussian models and GLMMs

$y = (y_1, ..., y_n)$ - outcome data over a set of $n$ locations

$y \sim p(y|g^{-1}(\eta), \theta)$ - observational model (likelihood)

$\eta = X\beta + f$ - linear predictor: fixed and random effects

$f \sim p(f|\theta)$ - random effect (RE)

# Latent Gaussian models and GLMMs

$y = (y_1, ..., y_n)$ - outcome data over a set of *n* locations

$y \sim p(y|g^{-1}(\eta), \theta)$ - observational model (likelihood)

$\eta = X\beta + f$ - linear predictor: fixed and random effects

$f \sim p(f|\theta)$ - random effect (RE)

$\theta \sim p(\theta)$ - hyperparameters

# Latent Gaussian models and GLMMs

| | |
|---|---|
| $y = (y_1, ..., y_n)$ | - outcome data over a set of $n$ locations |
| $y \sim p(y \| g^{-1}(\eta), \theta)$ | - observational model (likelihood) |
| $\eta = X\beta + f$ | - linear predictor: fixed and random effects |
| $f \sim p(f \| \theta)$ | - random effect (RE): Gaussian process |
| $\theta \sim p(\theta)$ | - hyperparameters |

# Outline

# Three types of spatial data

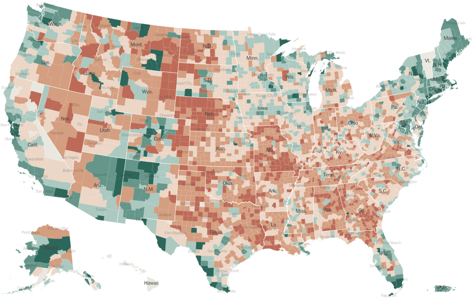
- Areal (or lattice)
- Geostatistical (or point-referenced)
- Point pattern

# Areal data

- Areal or lattice data arise when a fixed domain is partitioned into a finite number of *subregions* at which outcomes are aggregated

# Areal data

- Areal or lattice data arise when a fixed domain is partitioned into a finite number of *subregions* at which outcomes are aggregated




US vaccinations.

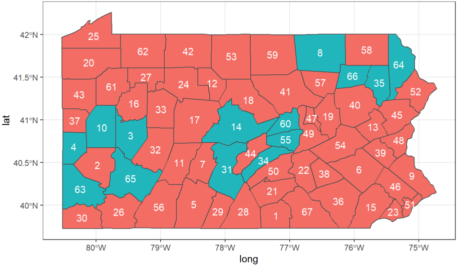
Credit: The New York Times

# Modelling areal data

- State-of-the-art models rely on "borrowing strength" from neighbours and use hierarchical Bayesian models to do so

# Modelling areal data

- State-of-the-art models rely on "borrowing strength" from neighbours and use hierarchical Bayesian models to do so




Neighbors of areas 2, 44 and 58 of Pennsylvania.

Credit: Moraga, "Geospatial Health Data: Modelling and Visualization with R-INLA and Shiny"

# Models of areal data

General form of the random effect prior (in terms of precision $Q$):

$$f \sim \text{MVN}(0, Q^{-1}).$$

# Models of areal data

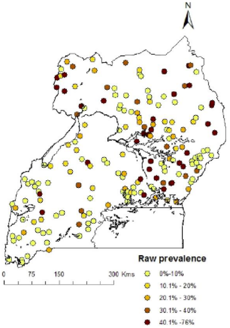
| | |
|---|---|
| $Q = \tau I$ | i.i.d. |
| $Q = \tau(D - \alpha A)$ | CAR: $A$ and $D$ are defined by the neighbourhood structure |
| $Q = \tau(D - A)$ | ICAR |
| $Q^{-1} = \tau_1^{-1} I + \tau_2^{-1}(D - A)^{-}$ | BYM |

# Geostatistical data

- Point-referenced (geostatistical) data represents measurements of a spatially continuous phenomenon at a set of *fixed* locations

# Geostatistical data

- Point-referenced (geostatistical) data represents measurements of a spatially continuous phenomenon at a set of *fixed* locations




Observed malaria prevalence at survey locations in Uganda, MIS 2014–15.

Credit: J Ssempiira

# Modelling geostatistical data

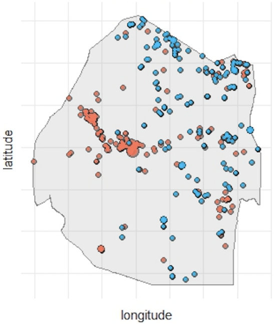
- Random effect at observed locations is usually modelled as a Gaussian process with some reasonable distance-dependent covariance kernel, such as RBF

$$k(x_i, x_j) = \tau \exp\left(-\frac{||x_i - x_j||^2}{2l^2}\right).$$

- Values at unobserved locations are predicted via kriging.

# Point pattern data

- Observed are the locations of events. Both the number of events and event locations are random.




Observed local (blue) and imported (red) malaria cases in Swaziland, 2015.

# Modelling point pattern data

- Point pattern data is modelled via point processes, such as the Poisson process.
- Models rely on the definition of the intensity function $\lambda(s), s \in D$.
- Log-Gaussian Cox process is a common model of spatial point patterns:

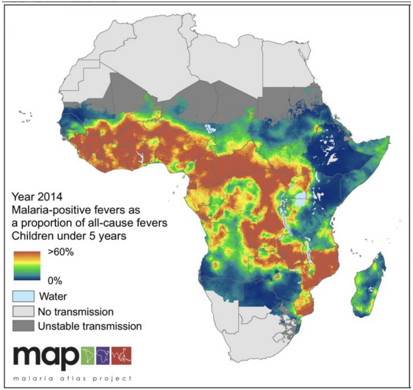
$$L(s_1, \ldots, s_n; \lambda(s)) = \exp(-\lambda(D)) \prod_{i=1}^{n} \lambda(s_i),$$

$$\lambda(D) = \int_D \lambda(s) ds,$$

$$\lambda(s) = \exp(X^T(s)\beta + f(s)).$$

# Scale of important public health problems

- Malaria prevalence in sub-Saharan Africa at 5x5 $km^2$ resolution.



# Probabilistic programming languages (PPLs) from a user's perspective

- PPLs allow users to specify probabilistic models and perform inference automatically.
- Users need to specify
  1. prior,
  2. likelihood,
- Inference is performed under the hood via powerful MCMC algorithms.

# Analyzing MCMC outputs

- Desired properties of MCMC chains:
  - Gelman-Rubin statistic ($\hat{R}$)
  - Effective sample size (ESS)

# Outline

# Goal

$$\underbrace{g(E[y|f_{\mathrm{GP}}])}_{\text{Linear predictor}} = \underbrace{X\beta}_{\text{Fixed effect}} + \underbrace{f_{\mathrm{GP}}}_{\text{Random effect}}$$

# Goal

$$g(E[y|f_{\text{GP}}]) = X\beta + f_{\text{GP}}$$

- Substitute evaluation of $f_{\text{GP}}$ (MVN) at inference stage with another variable which has better properties for MCMC inference.

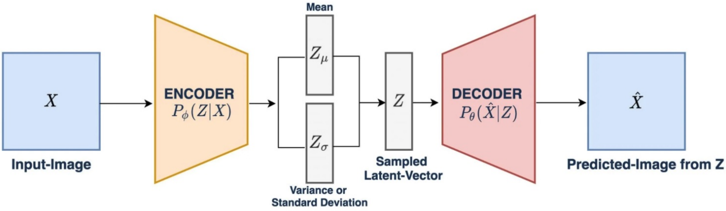
# Outline

# Idea: train VAE on GP priors

$$g(E[y|f_{\text{GP}}]) = X\beta + f_{\text{VAE}}$$

- Substitute evaluation of the GP (MVN) with the decoder of a trained VAE.

# Variational Autoencoder (VAE)




$$\boxed{\begin{array}{r} Z = \mu + \sigma \odot \epsilon \\ \epsilon \sim \mathcal{N}(0,1) \end{array}}$$

Credit: `https://learnopencv.com/variational-autoencoder-in-tensorflow/`

# Why should it work?

$z_n \sim N(0, I_n)$

$f_{\text{GP}} = L_\theta z_n$

Linear operation, but $\theta$ needs to be inferred.

$z_d \sim N(0, I_d), \quad d < n$

$f_{\text{VAE}} = \phi_w(z_d)$

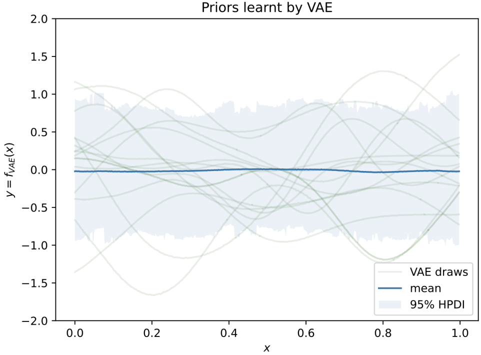
Non-linear operation, but deterministic transformation.

# 'PriorVAE' workflow

- Fix the spatial structure of interest (e.g. a grid or administrative boundaries).
- Use draws from a GP prior over the spatial structure as training data of a VAE.
- Use the trained decoder to substitute GP in a model used for inference.

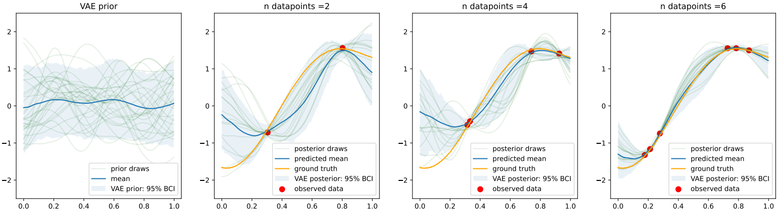
# Outline

# Example: 1d GP - learning priors



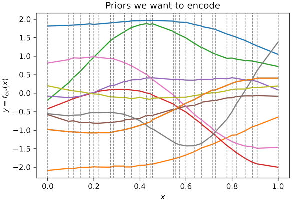

Learning one-dimensional GP priors on a regular grid:
$n = 400$, squared exponential kernel

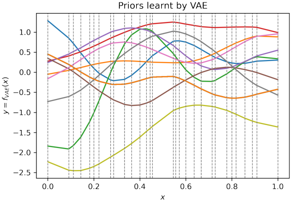
# Example: 1d GP - inference




Making inference using the learnt VAE prior on a regular grid, $n = 400$

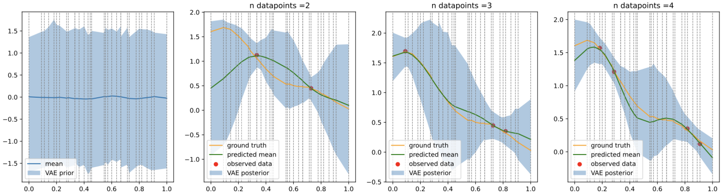
# Example: irregular grid - learning priors




Learning one-dimensional GP priors on an irregular grid, $n = 32$, squared exponential kernel

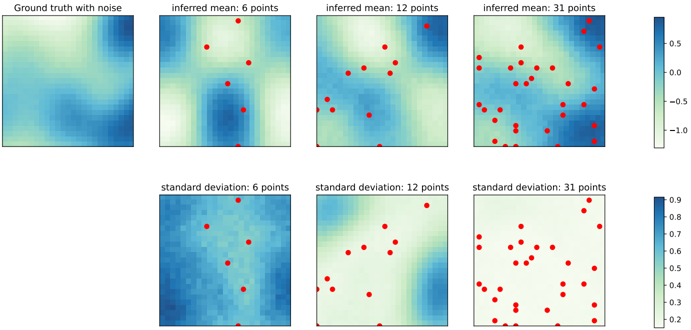
# Example: 1d GP - inference




Making inference using the learnt VAE prior on an irregular grid,

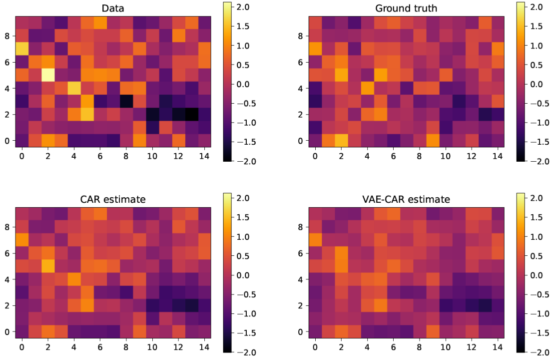
$$n = 32$$

# Example: 2d GP - inference




Making inference on a two-dimensional grid, $n = 25x25$

# Example: synthetic CAR




Partition a rectangle by subdividing it into $n_y = 10$ rows and $n_y = 15$ columns, creating 150 areas for CAR prior learning.

# Example: Scottish lip cancer data

- 56 counties of Scotland
- observed ($y$) and expected ($E$) number of lip cancer cases
- a covariate measuring the proportion of the population engaged in agriculture, fishing, or forestry $x$

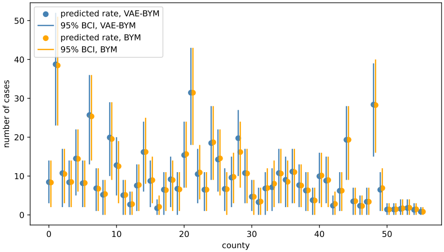
# Example: modelling Scottish lip cancer data

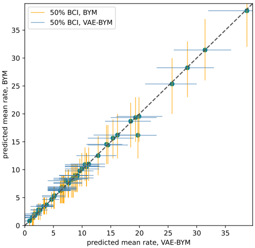
- BYM model:

$$
\begin{aligned}
y &\sim \text{Poisson}(\lambda), \\
\log(\lambda) &= \log(E) + b_0 + b_1 x + \underbrace{\phi_1 + \phi_2}_{\text{BYM prior}}, \\
\phi_1 &\sim N(0, \tau_1 I), \\
\phi_2 &\sim N(0, Q_2^-), \quad Q_2 = \tau_2(D - A).
\end{aligned}
$$

# Example: modelling Scottish lip cancer data

- VAE-BYM achieves goodness of fit very close to the one of BYM:



# Example: modelling Scottish lip cancer data

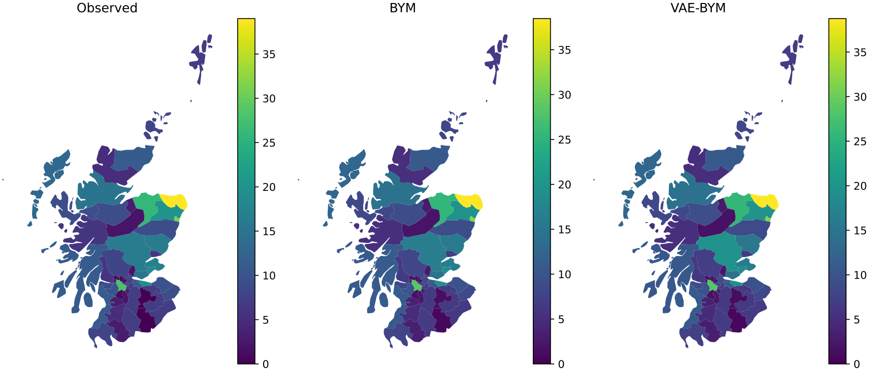
- Average ESS of the spatial effects

$$\text{BYM} : 150$$
$$\text{VAE-BYM} : 1030$$

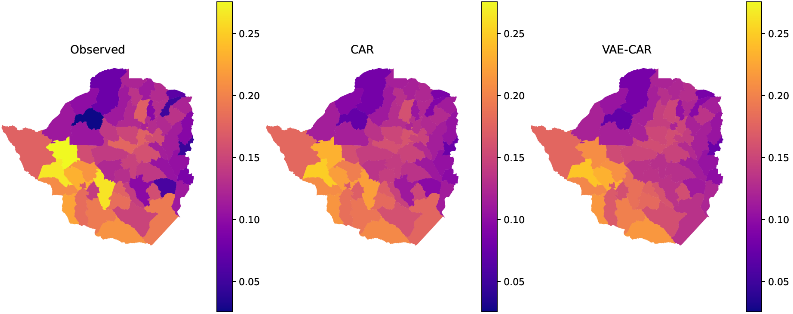
- MCMC elapsed time (s)

$$\text{BYM} : 402$$
$$\text{VAE-BYM} : 12$$

# Example: modelling Scottish lip cancer data

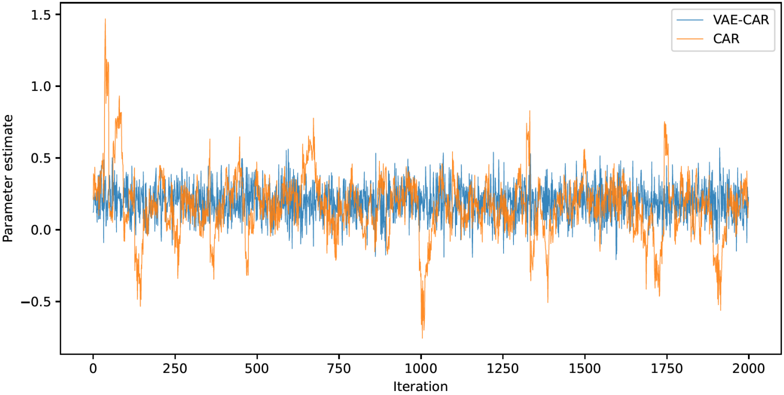


# Example: HIV prevalence in Zimbabwe

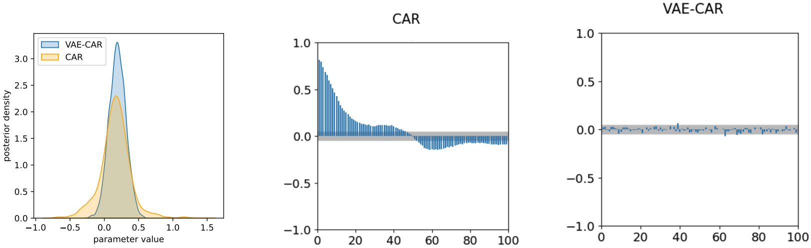



Results of MCMC inference on the observed number of HIV-positive individuals among observed individuals in Zimbabwe.

# Example: MCMC traceplot

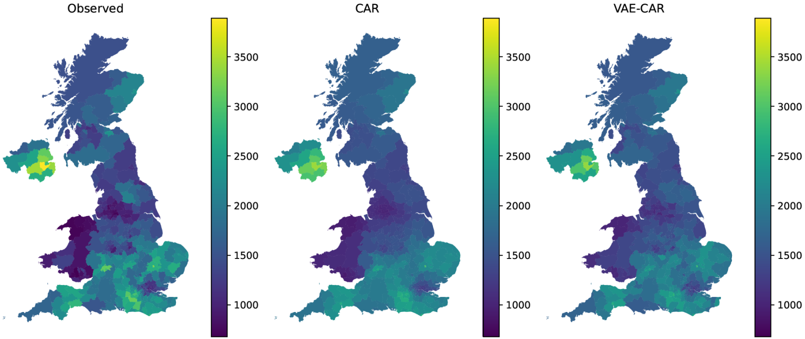


# Example: Zimbabwe, CAR model




Posterior density and autocorrelations of MCMC samples of a spatial random effect modelled via CAR and VAE-CAR.

# Example: UK LTLAs, Covid-19




MCMC inference on the projected COVID-19 incidence in the UK at the LTLA level.

# Outline

Latent Gaussian models and GLMMs

Three types of spatial data

What are we trying to solve?

Idea: learn GP priors with VAEs

Examples

Outro

# Summary

- What have we gained?
  - Substituted correlated structure with uncorrelated one,
  - Reduced dimensionality,
  - Higher ESS,
  - Lower computation time,
  - Unlimited data quantity,
  - Perfect data quality,
  - No need to store all of the data, only one batch.

# Summary

- What have we gained?
  - Substituted correlated structure with uncorrelated one,
  - Reduced dimensionality,
  - Higher ESS,
  - Lower computation time,
  - Unlimited data quantity,
  - Perfect data quality,
  - No need to store all of the data, only one batch.
- Future outlook:
  - Smarter losses (think "noise outsourcing lemma"),
  - Encode other distributions.

# Thank You!

elizaveta.p.semenova@gmail.com
@liza_p_semenova

# Example: 1d GP, empirical covariance matrices