Inference for Disease Mapping with Spatio-Temporal Gaussian Processes

Filipe P. de Farias[a] Francesco Corona[b] Michela Mulas[c]

Depto. de Engenharia de Teleinformática - Centro de Tecnologia, UFC

Introduction

Disease map is a collection of disease objects (residential locations of individuals or a summary measure or statistic for specified groups of individuals) in their geographical association.[3] With disease mapping we can:[1]

- Identify **disease patterns** in its spatial heterogeneity.
- Obtain clues as to the **disease aetiology** and highlight areas of **elevated risk**.

We define **risk** as how "**likely**" a disease can occur. Then, our objetive is to **infer** how spatially the risk is spatially distributed, given the **incidence** of the disease. Mapping the risk allows us to **develop prevention policies** and correlate with possible **risk factors**.

The work aims at studying a **Bayesian approach** to **spatio-temporal** data, more specifically the distribution of disease cases and its risk a country. The practical application is the inference of **Leprosy's risk** between the years of 2001 and 2019 in Brazil. The Leprosy, or Hansen's disease is a neglected tropical disease still present in Brazil.

Map Rasterization

We define a framework to allows us to control the **resolution** of the risk estimation. We associate to each municipality's its number of **notified cases**. We divide the map of the country in a grid (Figure 1). It is given the **position** $\mathbf{x}_n = [\text{latitude}_n, \text{longitude}_n]$ of the n-th cell, and there are N user defined cells.

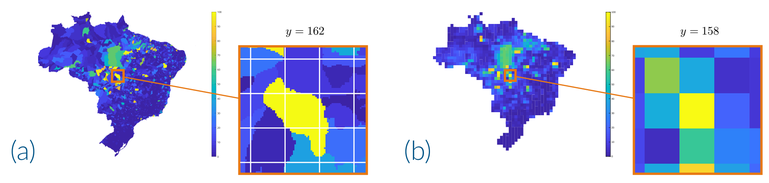


Figure 1. In (a) the municipalities with their respective case counts. In (b) the rasterization where each cell shows the correspondent average number of cases. Here was used the data for the year of 2017.

The n-th cell is characterized each the **observed number of cases** y_n and the **expected number of cases** e_n. The expected number of cases depends on its population p_n, the country's numbers of cases y_{tot} and population p_{tot} as in $e_n = p_n \cdot y_{\text{tot}}/p_{\text{tot}}$.[7] The risk is an unknown, or **latent** variable used for **estimating** the expected number of cases.

Bayesian Inference

The use of Bayesian inference allows us to obtain the **expected** (mean) and **variance** of the risk μ.

- We consider the process of the disease, specifically the **disease occurrence** y, to be **parametrised** by the risk μ (Figure 2).

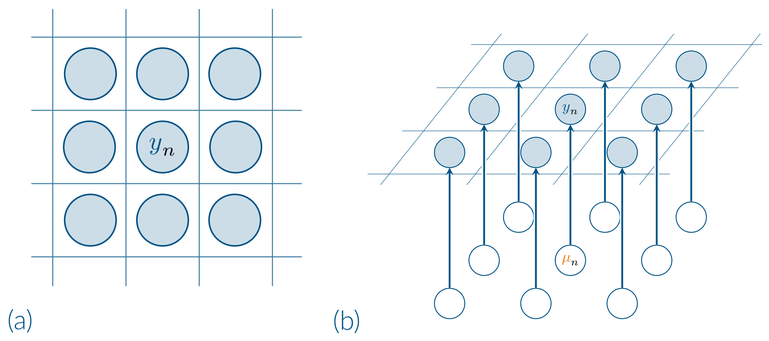


Figure 2. In (a) the graphical representation of the case counts y_n as a random variable in the n-th cell and its dependence on each cell's risks $[\mu_n]_{n=1}^N$ in (b).

From the **Bayesian** point of view, we can also represent the **uncertainty** about μ with **probabilities**. Then we combine the uncertainty about the data y with some **prior** about μ to update its uncertainty in the light of the observations.

In order to spatially relate the risks between cells we use **Gaussian Process** (GP) as prior for μ (Figure 3).

[a]Scientific Initiation at UFC. Email: filipepfarias@fisica.ufc.br.
[b]Email: francesco.corona@aalto.fi.
[c]Email: michela.mulas@ufc.br.

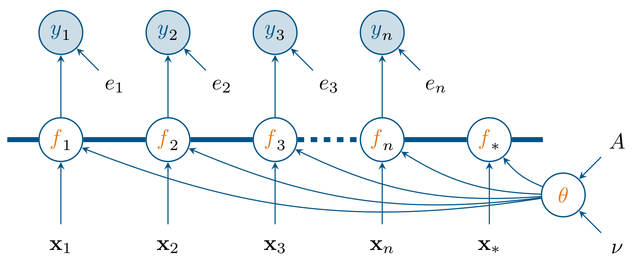


Figure 3. The graphical model for the spatial risk as a Gaussian Process. The thicker line represents the Gaussian Field,[5] in which each $\mathbf{f} = [f_n]_{n=1}^N$ is a marginalisation of the latent process f at the n-th cell.

The probabilistic model for this problem is

$$\begin{cases} y_1, y_2, \ldots, y_n | \mu \sim \prod_{n=1}^{N} Poisson(e_n \mu_n) & \text{(1a)} \\ f_n = \log(\mu_n) & \text{(1b)} \\ f(\mathbf{x}) | \theta \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}' | \theta)\right) & \text{(1c)} \\ [\theta]_p \sim \text{half-}t(\nu, s^2). & \text{(1d)} \end{cases}$$

The **observation model** (1a) comprises the **independent** observations $\mathbf{y} = [y_n]_{n=1}^N$ of the process, which will be assumed **Poisson distributed**, as we're modelling a count process. The GP **prior model** in (1c) represents some previous knowledge about the risk μ, it is parametrised by a mean function $m(\mathbf{x}) = 0$ and a covariance function[5] $k(\mathbf{x}, \mathbf{x}'|\theta) = k_{\text{sexp}}(\mathbf{x}, \mathbf{x}'|\theta) + k_{\text{matern}}(\mathbf{x}, \mathbf{x}'|\theta)$. Assuming this mean function is equivalent to initially assume the same risk for all the cells. Assuming this covariance function accounts for long and short range similarities between cells. As the risk is positive definite, we transform it as in (1b).

The variable θ denotes the **unknown hyperprior** in (1d). It's responsible for parametrizing the GP **covariance function**. We model θ as half-Student's t distribution, a **weakly informative hyperprior** with **known hyperparameters** (ν, s^2).

The log **risk posterior** $\mathbf{f}$ is evaluated as

$$\begin{aligned} p(\mathbf{f} \mid \mathbf{y}) &= \int p(\mathbf{f}, \theta \mid \mathbf{y}) d\theta \\ &= \frac{1}{Z_p} \int \overbrace{p(\mathbf{y} \mid \mathbf{f})}^{\text{Poisson Obs. Model}} \underbrace{p(\mathbf{f} \mid \mathbf{X}, \theta)}_{\text{GP Prior Model}} \overbrace{p(\theta)}^{\text{half-}t\text{ Hyperprior}} d\theta, \end{aligned}$$

where $\mathbf{X} = [\mathbf{x}]_{n=1}^N$ is the collection of cell's positions, and Z_p the **normalisation constant**

$$Z_p = \iint p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, \theta)\, p(\theta) d\theta d\mathbf{f}.$$

Bayesian Filtering

The objective is to update the posterior $p(\mathbf{f}|\mathbf{y})$ at each step t. For the task we use data using **Bayesian filtering**.[2]

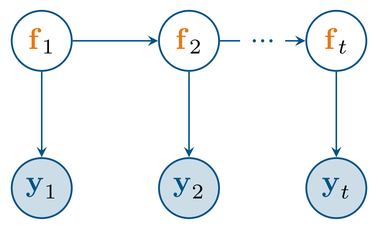


Figure 4. Graphical model of the Markov Chain for the iterative update of the latent variable $\mathbf{f}$.

We denote by $\mathbf{f}_{1:t} = [\mathbf{f}_t]_{t=1}^T$ the time-series latent risk. We assume this process being **markovian**, i.e. the probability of the the risk $\mathbf{f}_t$ depends only on $\mathbf{f}_{t-1}$, $p(\mathbf{f}_t \mid \mathbf{f}_{t-1})$. Similarly the observation $\mathbf{y}_t$ only depends on $\mathbf{f}_t$, $p(\mathbf{y}_t \mid \mathbf{f}_t)$. Using **Bayes' Rule** to derive the **prediction** and **update** steps of the **Bayesian filtering equations**.[6] The predictive distribution of $\mathbf{f}_t$, given a dynamic model $p(\mathbf{f}_t|\mathbf{f}_{t-1})$, is computed by the **Chapman-Kolmogorov equation**:

$$\begin{aligned} p(\mathbf{f}_t \mid \mathbf{y}_{1:t-1}) &= \int p(\mathbf{f}_t \mid \mathbf{f}_{t-1})\, p(\mathbf{f}_{t-1} \mid \mathbf{y}_{1:t-1})\, d\mathbf{f}_{t-1} \\ &= \frac{p(\mathbf{y}_t \mid \mathbf{f}_t)\, p(\mathbf{f}_t \mid \mathbf{y}_{1:t-1})}{\int p(\mathbf{y}_t \mid \mathbf{f}_t)\, p(\mathbf{f}_t \mid \mathbf{y}_{1:t-1}) d\mathbf{f}_t}. \end{aligned}$$

Joining the prediction and update steps we get

$$p(\mathbf{f}_t \mid \mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t \mid \mathbf{f}_t) \int p(\mathbf{f}_t \mid \mathbf{f}_{t-1})\, p(\mathbf{f}_{t-1} \mid \mathbf{y}_{1:t-1})\, d\mathbf{f}_{t-1}}{\int p(\mathbf{y}_t \mid \mathbf{f}_t) \int p(\mathbf{f}_t \mid \mathbf{f}_{t-1})\, p(\mathbf{f}_{t-1} \mid \mathbf{y}_{1:t-1})\, d\mathbf{f}_{t-1}\, d\mathbf{f}_t}.$$

As this does not express a **conjugated distribution**, then we can not obtain its summaries like mean and variance explicitly. To overcame this limitations we apply Laplace approximation to obtain $p(\mathbf{f}_t|\mathbf{y}_{1:t})$ summaries.

Conducting the Inference

To conduct the inference we use **Laplace approximation**.[5]

$$p(\mathbf{f}_t \mid \mathbf{y}_{1:t}) \approx q(\mathbf{f}_t \mid \mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{f}_t \mid \hat{\mathbf{f}}_t, \mathbf{C}_t),$$

where the posterior mode is $\hat{\mathbf{f}}_t = \arg\max_{\mathbf{f}_t} p(\mathbf{f}_t|\mathbf{y}_{1:t})$, and the posterior curvature is $\mathbf{C}_t^{-1} = -\nabla\nabla \ln p(\mathbf{f}_t|\mathbf{y}_{1:t})|_{\mathbf{f}_t = \hat{\mathbf{f}}_t}$.

We have that the logarithm of the posterior is

$$\begin{aligned} \Psi(\mathbf{f}_t) = \log p(\mathbf{y}_t|\mathbf{f}_t) - \tfrac{1}{2}(\mathbf{f}_t - \boldsymbol{\mu}_t^-)^\top (\mathbf{C}_t^-)^{-1}(\mathbf{f}_t - \boldsymbol{\mu}_t^-) \\ - \tfrac{1}{2}\log|\mathbf{C}_t^-| - \tfrac{n}{2}\log 2\pi, \end{aligned}$$

here $\mathbf{C}_t^- = \mathbf{A}\mathbf{C}_{t-1}\mathbf{A}^\top + \mathbf{Q}$ and $\boldsymbol{\mu}_t^- = \mathbf{A}\boldsymbol{\mu}_{t-1}$. $\mathbf{A}$ is the **dynamic matrix**. Differentiating Ψ w.r.t. $\mathbf{f}_t$ we get

$$\begin{aligned} \nabla\Psi(\mathbf{f}_t) &= \nabla \log p(\mathbf{y}_t \mid \mathbf{f}_t) - (\mathbf{C}_t^-)^{-1}(\mathbf{f}_t - \boldsymbol{\mu}_t^-), \\ \nabla\nabla\Psi(\mathbf{f}_t) &= -\nabla\nabla \log p(\mathbf{y}_t \mid \mathbf{f}_t) - (\mathbf{C}_t^-)^{-1} = -\mathbf{W} - (\mathbf{C}_t^-)^{-1}. \end{aligned}$$

To obtain $\hat{\mathbf{f}}_t$, we maximize Ψ,[5] which implies

$$\nabla\Psi(\hat{\mathbf{f}}_t) = \mathbf{0} \Longrightarrow \hat{\mathbf{f}}_t = \mathbf{C}_t^-(\nabla \log p(\mathbf{y}_t \mid \hat{\mathbf{f}}_t)) + \boldsymbol{\mu}_t^-.$$

To solve for $\hat{\mathbf{f}}_t$, we use Newton's method

$$\begin{aligned} \mathbf{f}_t^{\text{new}} &= \mathbf{f}_t - (\nabla\nabla\Psi)^{-1}\nabla\Psi \\ &= \mathbf{f}_t + [\mathbf{W} + (\mathbf{C}_t^-)^{-1}]^{-1}[-\nabla \log p(\mathbf{y}_t \mid \mathbf{f}_t) \\ &\quad - (\mathbf{C}_t^-)^{-1}(\mathbf{f}_t - \boldsymbol{\mu}_t^-)] \\ &= [\mathbf{W} + (\mathbf{C}_t^-)^{-1}]^{-1}[\mathbf{W}\mathbf{f}_t + \nabla \log p(\mathbf{y}_t \mid \mathbf{f}_t) + (\mathbf{C}_t^-)^{-1}\boldsymbol{\mu}_t^-]. \end{aligned}$$

Discussion

The annual posterior of the risk of incidence of Leprosy is attached. We can see that the North, Northeast and Midwest regions of Brazil are characterized by higher expected values of the risk $\mathbb{E}[\mu]$, compared to the South and the Southeast. This is true over the entire observation period, a result that is expected.[4]

We can also note that applying Bayesian filtering reduces the uncertainty $\mathbb{V}\text{ar}[\mu]$ of the posterior.

The computational cost of this solution is constrained by the well known GP drawback of matrix inversion. In our case this corresponds to the inversion of a $N \times N$ matrix, whose cost increases with the resolution of the grid.

The Bayesian inference with GP leads to valuable and interpretable results. It's important to also note that the evolution of the risk in time is not accounted for, i.e. in the calculations that the dynamic matrix $\mathbf{A}$ is assumed to be time-invariant and equal to the identity. A more accurate characterization of the dynamics is the objective of the future works.

References

[1] Nicky Best, Sylvia Richardson, and Andrew Thomson.
A comparison of bayesian spatial models for disease mapping.
Statistical Methods in Medical Research, 14(1):35–59, 2005.

[2] Michael C. et al Burkhart.
The discriminative kalman filter for bayesian filtering with nonlinear and nongaussian observation models.
Neural Computation, 32(5):969–1017, 2020.

[3] A.B. Lawson, A. Lawson, and F.L.R. Williams.
An Introductory Guide to Disease Mapping.
Pharmaceutical Medicine. Wiley, 2001.

[4] Francisco Rogerlândio et al Martins-Melo.
Leprosy-related mortality in Brazil: a neglected condition of a neglected disease.
Transactions of The Royal Society of Tropical Medicine and Hygiene,
109(10):643–652, 09 2015.

[5] C.E. Rasmussen and C.K.I. Williams.
Gaussian Processes for Machine Learning.
Adaptive computation and machine learning. MIT Press, 2006.

[6] Simo Särkkä.
Bayesian filtering and smoothing, volume 3.
Cambridge University Press, 2013.

[7] Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari.
Approximate inference for disease mapping with sparse gaussian processes.
Statistics in medicine, 29(15):1580–1607, 2010.

[8] Ke Alexander Wang, Geoff Pleiss, Jacob R. Gardner, Stephen Tyree, Kilian Q. Weinberger, and Andrew Gordon Wilson.
Exact gaussian processes on a million data points, 2019.

Inference for Disease Mapping with Spatio-Temporal Gaussian Processes

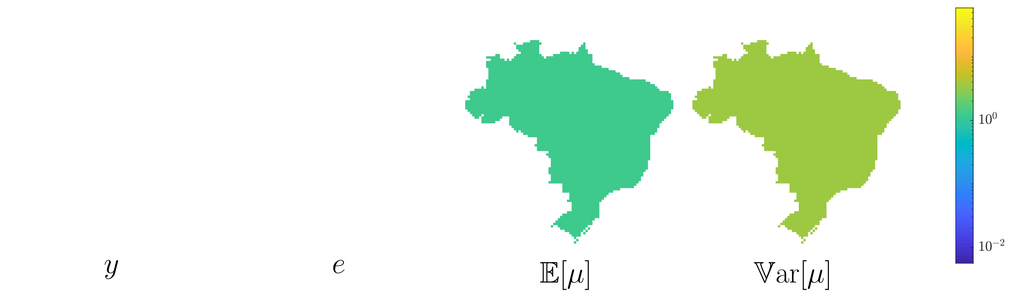
Filipe P. de Farias[a] Francesco Corona[b] Michela Mulas[c]

Depto. de Engenharia de Teleinformática - Centro de Tecnologia, UFC

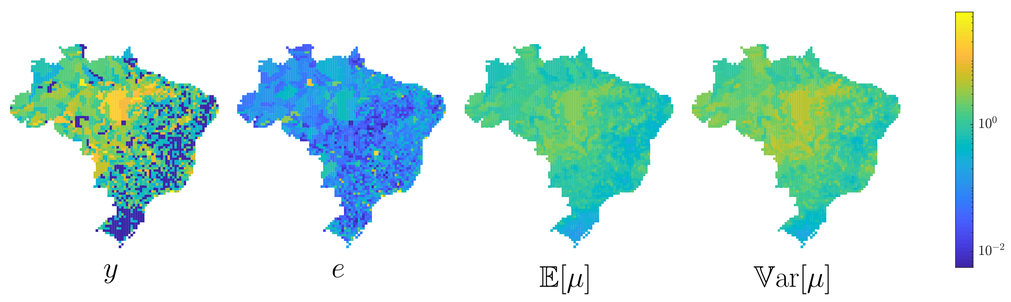
ENCONTROS UNIVERSITÁRIOS

UNIVERSIDADE FEDERAL DO CEARÁ

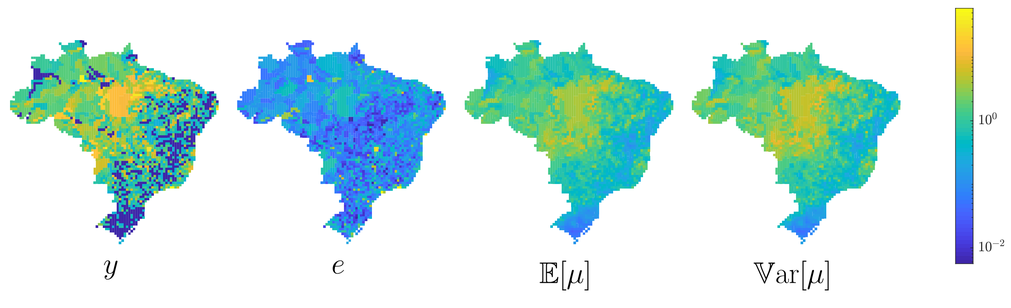
Prior



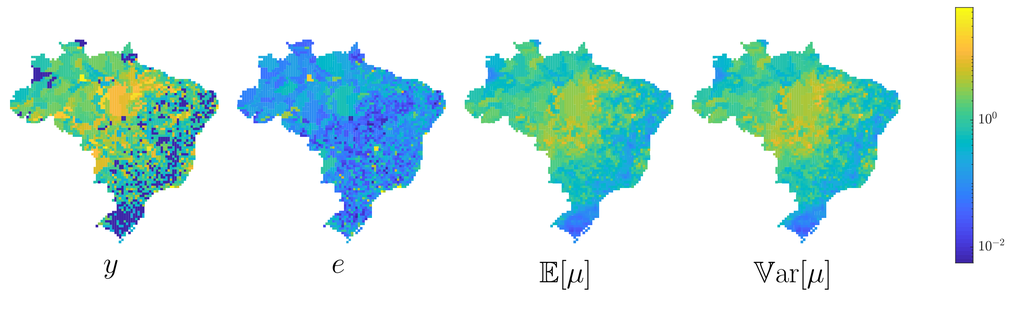
2001



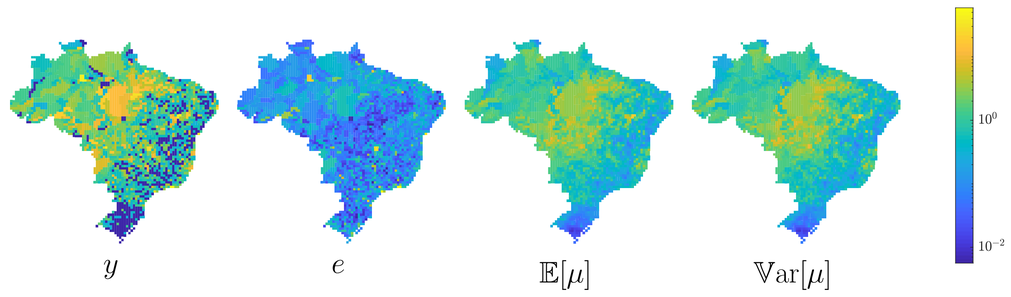
2002



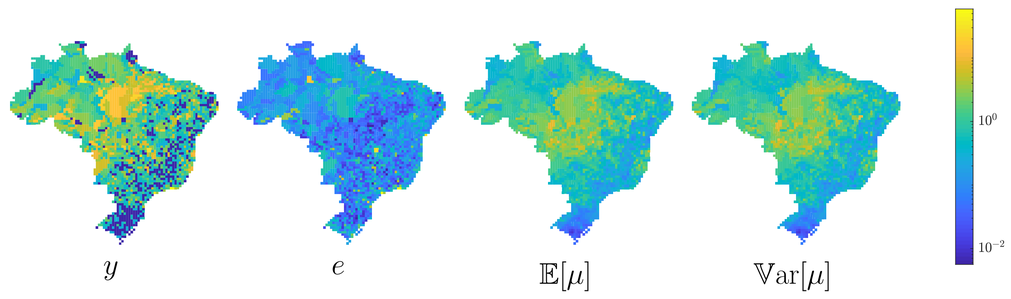
2003



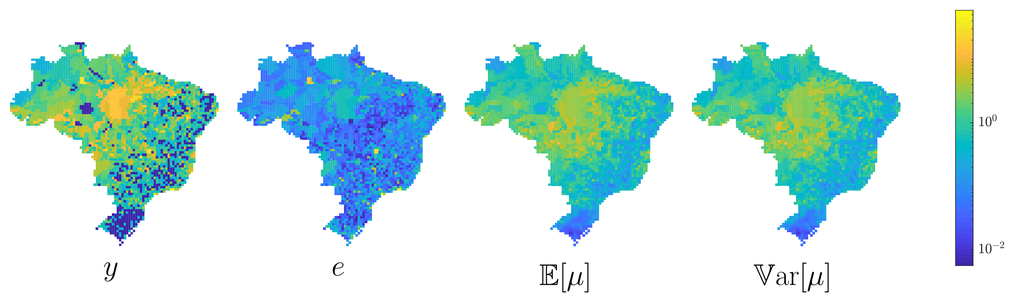
2004



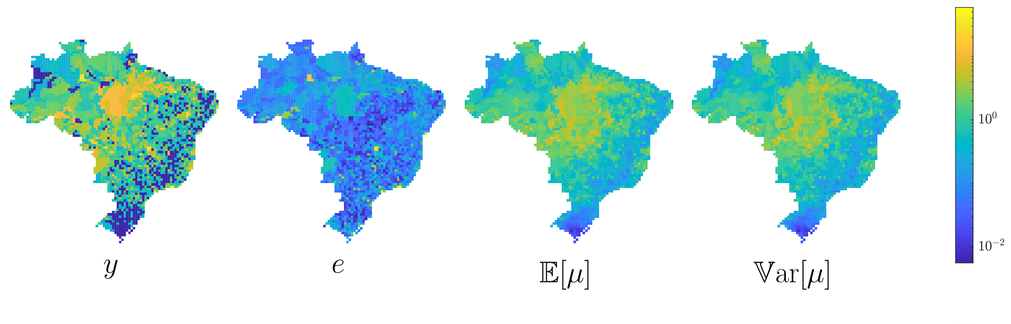
2005



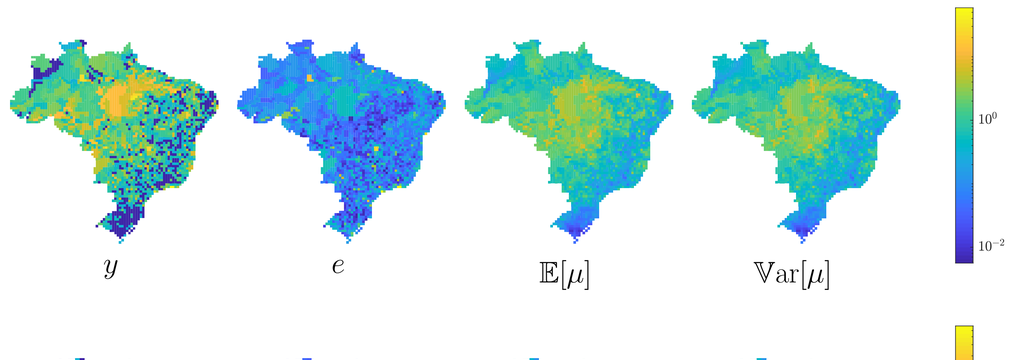
2006



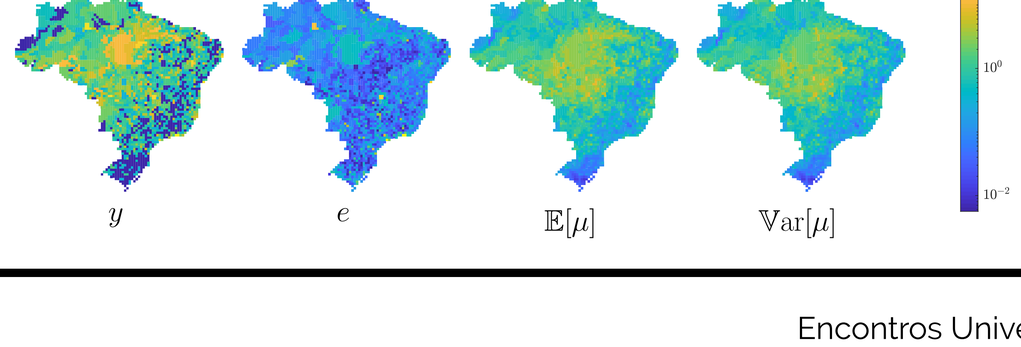
2007



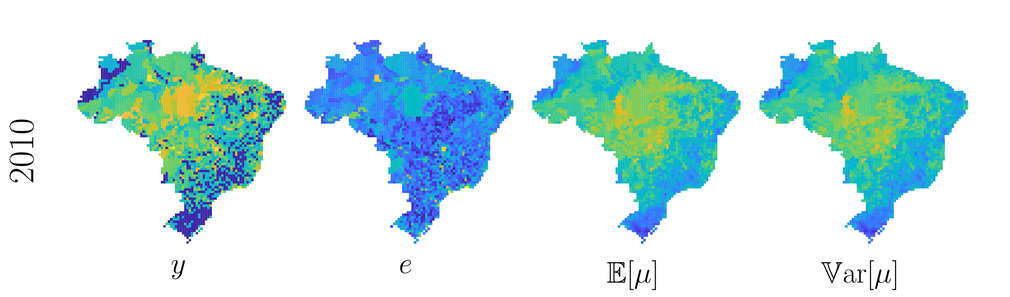
2008



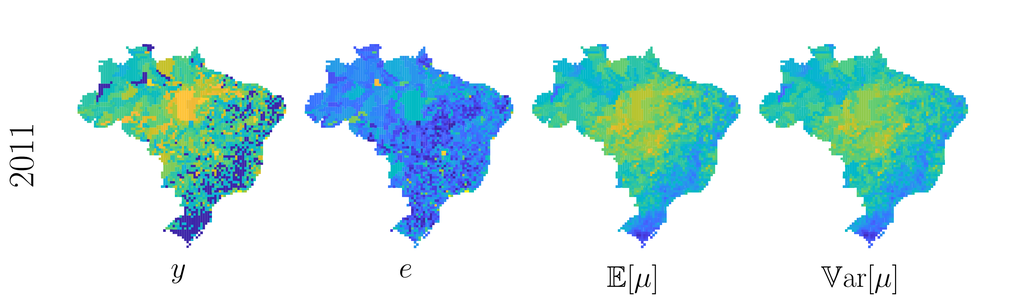
2009



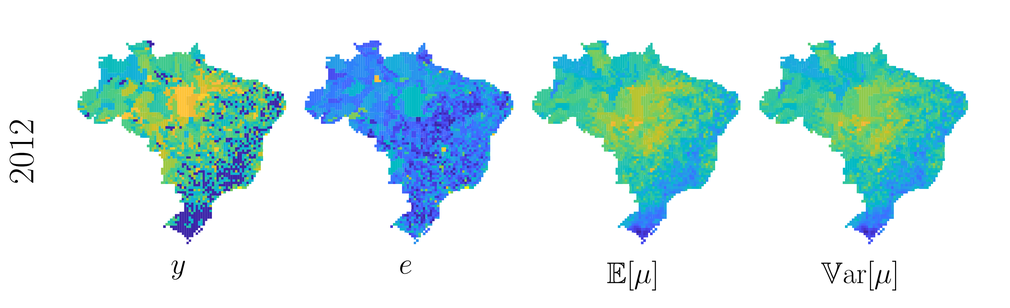
2010



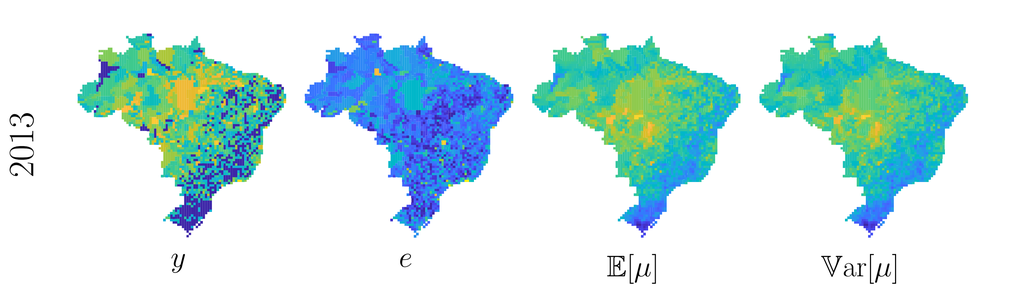
2011



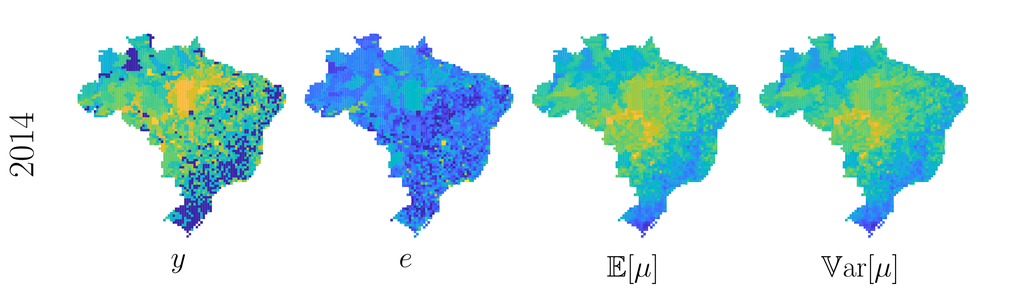
2012



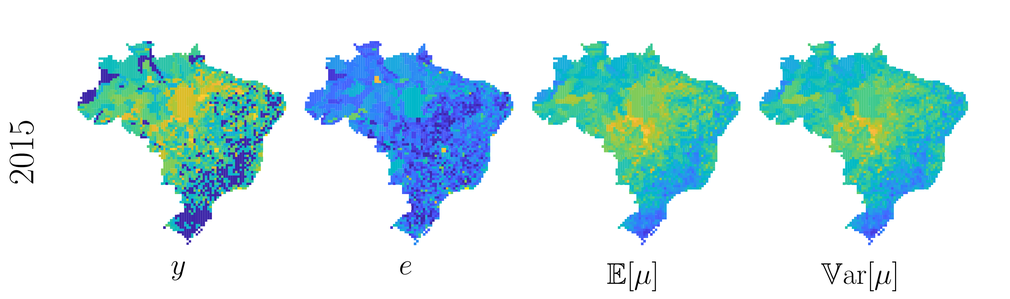
2013



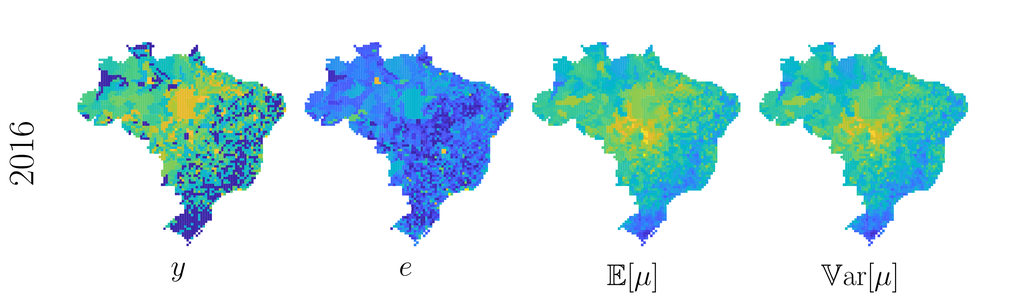
2014



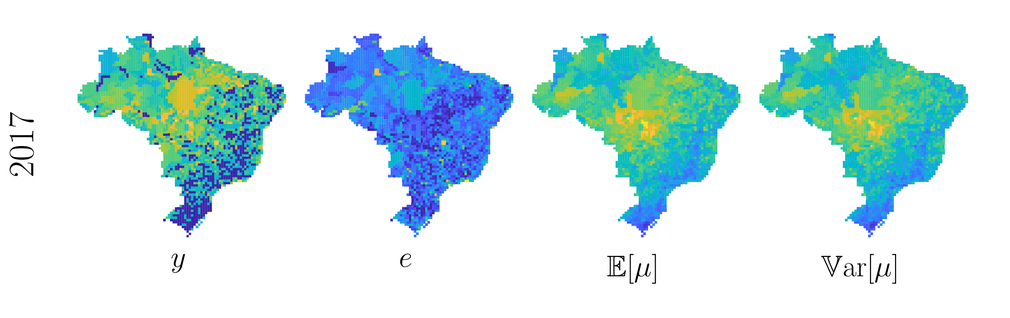
2015



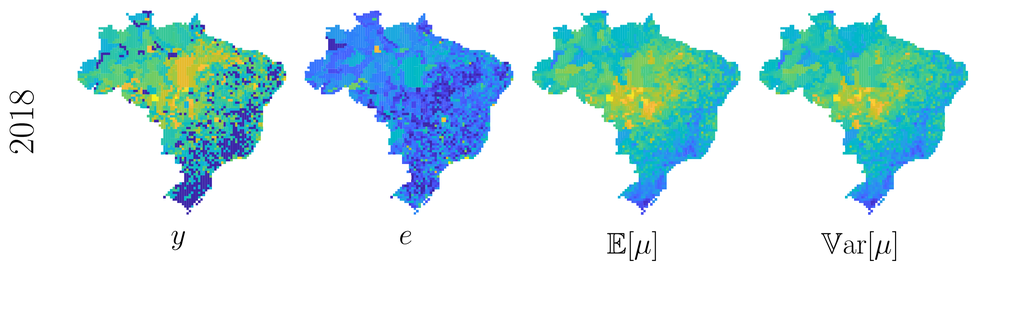
2016



2017



2018



2019