An unified framework for point-level, areal, and mixed spatial data: The Hausdorff-Gaussian Process

Lucas da Cunha Godoy

Department of Statistics, University of Connecticut

June 28, 2022

Join work with

Marcos Prates

Jun Yan

Outline

Introduction

- Taking into account spatial dependence possibly present in data is a foremost aspect of spatial statistics.
- The way researchers usually deal with the spatial structures varies with the nature (or geometry) of the observed spatial data.
- In this work, we propose the use of the Hausdorff distance (Shephard and Webster 1965 ; see pg. 280, Grünbaum and Shephard 1969) to unify the theory used for areal and geostatistical data.

- The Hausdorff distance is capable of taking into account shapes and sizes of regions and, when applied to "pairs of coordinates" it can be reduced to the Euclidean distance.
- The central idea is extending the methodology used in geostatistics to more general geometries (i.e., areal data).

A Unified Framework

- Suppose we are working with a study region, typical a map, $D \subset \mathbb{R}^{\mathrm{d}}$.
- Typically $d = 2$ or $d = 3$.
- Define $\mathcal{B}_D$ as the class of closed subsets of $D \subset \mathbb{R}^{\mathrm{d}}$.
- Then, a slightly more general definition of a random field is as follows

$$\{S(B) : B \in \mathcal{B}_D\}.$$

- The definition $\{S(\mathbf{s}) : \mathbf{s} \in D\}$, usually used in geostatistics, is a special case of the definition given in the last slide.
- Pairs of coordinates (point-level data) can be thought of as a single vertex (or a "collapsed") polygon.
- Lastly, and even more importantly, such definition has the ability to handle areal data.
- Although not a necessary condition, we will assume the random fields $S(B)$ to be stationary and isotropic in this work.

Distance Between Polygons

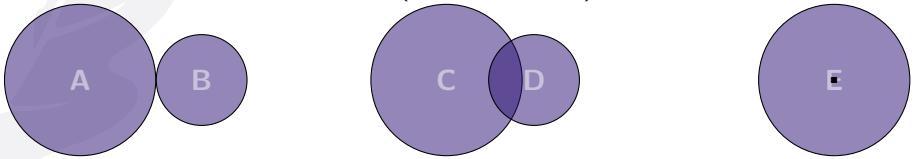
- Distance between borders (**not a metric**):

$$d_b(B_i, B_j) = \inf_{\mathbf{p} \in B_i, \mathbf{q} \in B_j} \|\mathbf{p} - \mathbf{q}\|$$

- "Integral" distance (Gotway and Young 2007)

$$d_I(A, B) = \int_{A \times B} \|\mathbf{p} - \mathbf{q}\| \, \mathrm{d}\mathbf{q} \, \mathrm{d}\mathbf{p}$$

- Distance between the centroids (**not a metric.**).



(a) (b) (c)

Hausdorff distance

- The Hausdorff distance tries to quantify the worst case scenario cost to travel from a set A towards another set B and vice-versa.
- Formally, we have

$$H_{\mathrm{d}}(A,B) = \max\left\{ \sup_{\mathbf{p}\in A}\left\{ \inf_{\mathbf{q}\in B} \mathrm{d}(\mathbf{p},\mathbf{q}) \right\}, \sup_{\mathbf{q}\in B}\left\{ \inf_{\mathbf{p}\in A} \mathrm{d}(\mathbf{p},\mathbf{q}) \right\} \right\},$$

where d is any distance metric, e.g. Manhattan distance, the geodesic distance, and the Euclidean distance.

- The distance $H_{\mathrm{d}}(A,B)$ **is a metric** (Shephard and Webster 1965)!

Hausdorff-Gaussian Processes

- We propose a new class Gaussian Process based on the more general definition of random field (e.g. $S(B)$) given in the previous slides, the *Hausdorff-Gaussian Process* (HGP).
- The process is capable of quantify the spatial correlation for a variable observed at general spatial geometries by the use of the Hausdorff distance.
- To exemplify, suppose we are working with a realization of an isotropic HGP with a positive-definite correlation function $\rho(\cdot)$. The spatial correlation between two sites B_i and B_j may be quantified by

$$\rho(B_i, B_j; \theta) = \rho(H(B_i, B_j); \theta),$$

  where θ represents the parameters controlling the spatial structure of the given correlation model.
- This definition is simple, flexible, and extremely useful. From now on, although not a theoretical restriction, we are going to assume the HGPs to be isotropic.

Why using HGP?

- Conditional autoregressive models (Besag 1974, CAR) and its variants (Besag, York, and Mollié 1991; Leroux, Lei, and Breslow 2000; Assunção and Krainski 2009) are often preferred when working with areal data.
- The CAR-like models are based on the assumption that the value of a variable observed at a specific areal unit is driven by the average of its neighbors.
- These models are computationally efficient (Rue and Held 2005; Rue, Martino, and Chopin 2009), especially if compared to geostatistical models. However, they suffer from the following problems
  1. They are not suitable for predictions.
  2. There exists a lack of interpretability of spatial parameters (Assunção and Krainski 2009; Datta et al. 2019).
  3. The spatial dependence does not take into account the size (nor the shape) of the areas.

- An alternative to the CAR models (Gelfand, Zhu, and Carlin 2001; Gotway and Young 2007; Moraga et al. 2017; White, Gelfand, and Utlaut 2017) is to consider $Y(B) = \frac{1}{|B|}\int_B Y(\mathbf{s})\,\mathrm{d}\mathbf{s}$. We will call this models areal averaged geostatistics (AAG).
- These models have been criticized for applications on which the nature of the observed variables contradicts the hypothesis regarding the aggregation of a continuous underlying process (Gelfand, Zhu, and Carlin 2001; White, Gelfand, and Utlaut 2017)
- This approach relies on Monte Carlo (or numeric) methods to evaluate numeric integrals over arbitrary polygons. Thus, it does not scale and quickly becomes computationally prohibitive.
- In addition, the numerical integrals can yield to biases that are hard (if not impossible) to be quantified (Gonçalves and Gamerman 2018).

In summary, the HGP

1. Unifies the models for geostatistics and areal data.
2. Allows for predictions based on models for areal data by construction.
3. Is suitable both for spatial misalignment and data fusion problems.
4. Scales better than the AAG models.
5. Its implementation and interpretation are trivial!

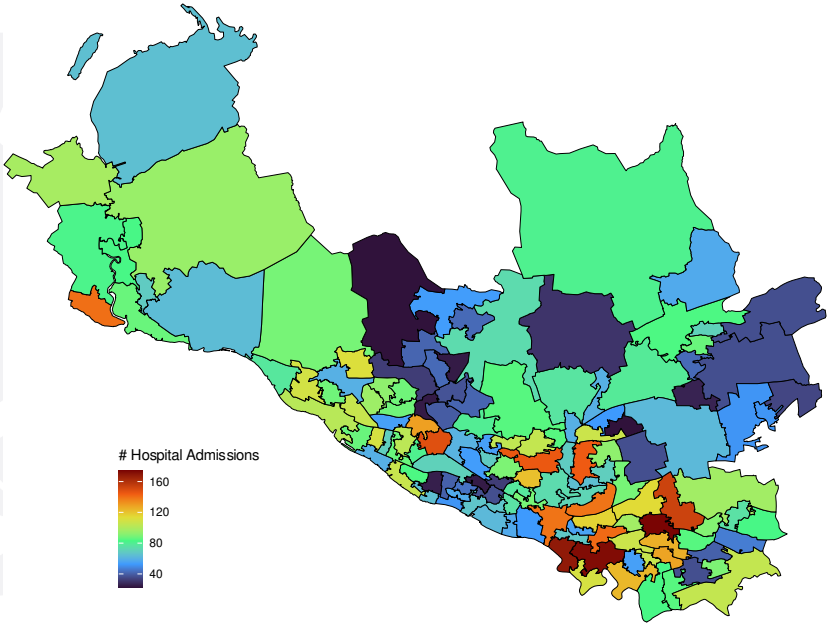
Data Analysis: Scotland Respiratory Cancer Data

- The dataset is freely available from the `CARBayes` `R` package (Lee 2013).
- It contains information about the number of hospital admissions by respiratory cancer in the Great Glasgow in Scotland in 134 regions called intermediate geographies (IG).
- The expected number of admissions based on age and sex standardized rates for the whole of Scotland is also available.
- The absolute difference in the percentage of people who are defined to be income deprived is an important predictor available.

The goals of the analysis are:

1. to provide good estimates, controled by income deprivation, for the standardized incidence ratio (SIR).
2. to compare the results of the proposed methodology to the Leroux (Leroux, Lei, and Breslow 2000) and ICAR (Besag, York, and Mollié 1991) models.

We will show that using the HGP as a random effect provide comparable model fit as those from the Leroux model with the advantages that the spatial parameter is interpretable and we may calculate the predicted SIR for other spatial aggregations of the same city without any extra effort.


Hospital Admissions
160
120
80
40

The Model

The likelihood considered for the three models is the same except for the random effect and is given by

$$Y_i \mid \lambda_i \sim \text{Poisson}(\lambda_i E_i)$$
$$\log(\lambda_i) = \alpha + \beta x_i + s_i,$$

where $s_i = s(B_i)$, x_i is the the absolute difference in the percentage of people who are defined to be income deprived at the region i, E_i is the expected number of cases for the i-th region, Y_i is the number of admission for the region i, and λ_i is the respective expected standardized incidence ratio.

Priors

Table 1: Priors for the models associated with the different models.

Model	Parameter	Prior
ICAR	τ	$\mathcal{G}(1,1)$
Leroux	τ	$\mathcal{G}(1,1)$
	ψ	$\mathcal{U}(0,1)$
HGP	τ	$\mathcal{G}(1,1)$
	ϕ	$\mathcal{U}(a_\phi, b_\phi)$
ALL	α	$\mathcal{N}(0, 100^2)$
	β	$\mathcal{N}(0, 100^2)$

Model comparison using WAIC

Table 2: WAIC for the HGP, Leroux, and ICAR random effects.

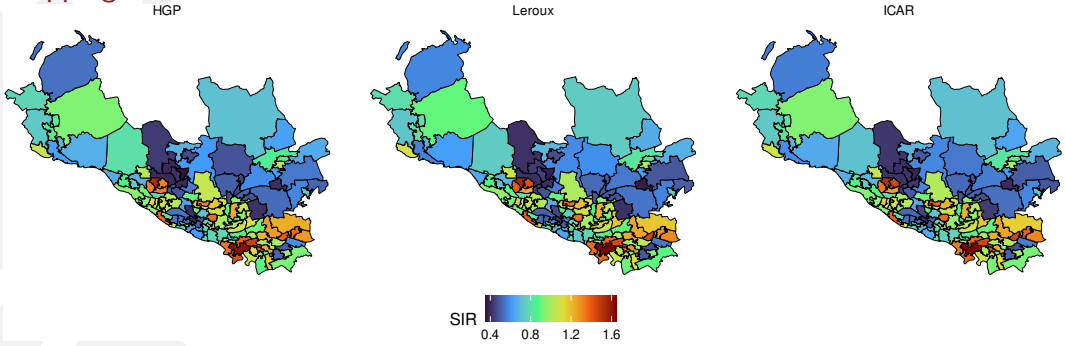
HGP	Leroux	ICAR
1029.782	1018.561	1039.583

Parameters estimation

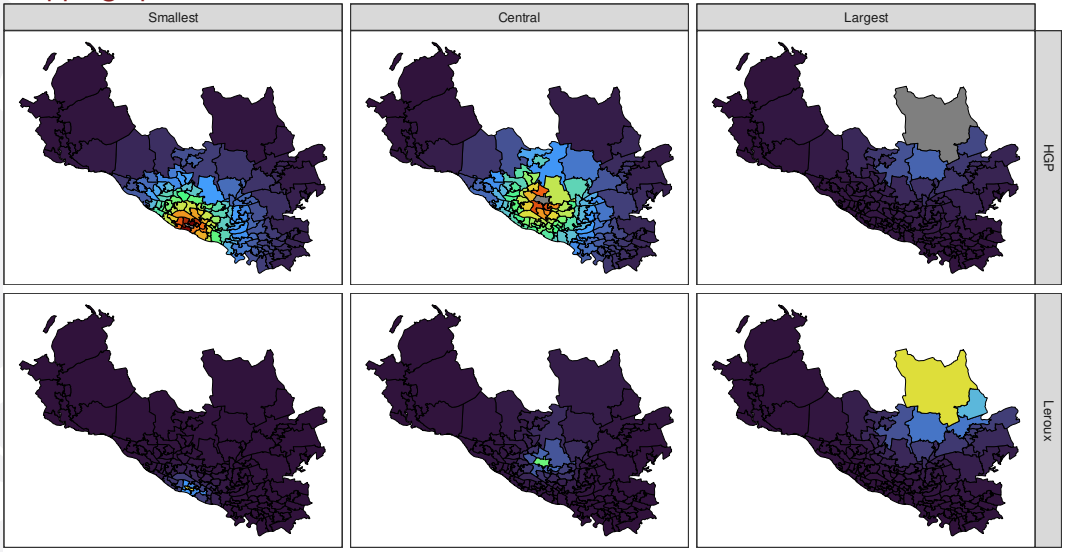
Table 3: Posterior expected value and highest posterior density interval for the parameters associated with the three different models applied to the data.

θ	Model	$\mathbb{E}[\theta \mid \cdot]$	95 % HPD
α	HGP	-0.222	[-0.242; -0.201]
	ICAR	-0.749	[-0.832; -0.665]
	Leroux	-0.778	[-0.835; -0.724]
β	HGP	0.024	[0.021; 0.027]
	ICAR	0.024	[0.020; 0.028]
	Leroux	0.025	[0.023; 0.027]
ϕ	HGP	3.700	[0.639; 7.481]
ψ	Leroux	0.540	[0.225; 0.861]
τ	HGP	11.773	[5.177; 18.552]
	ICAR	6.263	[4.282; 8.489]
	Leroux	9.029	[5.573; 13.259]

Mapping SIR



Mapping spatial correlations



Dicussion

- We introduce an unified framework (based on the Hausdorff distance) for spatial data that is flexible enough to deal with areal and geostatistical data.
- We have shown how useful and flexible it is for analyzing areal data.
- Although appropriate for polygons, the Hausdorff distance is not perfect. As it is based on worst case analysis (Paumard 1997), when computing the distance between two polygons, it can be heavily affect by an "outlier" vertex.
- The methodology also inherits the "big n" problem from geostatistical methods.
- Although most methodologies used in geostatistics are naturally extended to HGPs in any spatial seeting, it may be hard, for example, to deal and interpret with anisotropy for areal data.

Thank you

- lcgodoy.me
- lcgodoy
- ldcgodoy

References I

Assunção, Renato, and Elias Krainski. 2009. "Neighborhood Dependence in Bayesian Spatial Models." *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 51 (5): 851–69.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.

Datta, Abhirup, Sudipto Banerjee, James S Hodges, and Leiwen Gao. 2019. "Spatial Disease Mapping Using Directed Acyclic Graph Auto-Regressive (Dagar) Models." *Bayesian Analysis* 14 (4): 1221.

Gelfand, Alan E, Li Zhu, and Bradley P Carlin. 2001. "On the Change of Support Problem for Spatio-Temporal Data." *Biostatistics* 2 (1): 31–45.

Gonçalves, Flávio B, and Dani Gamerman. 2018. "Exact Bayesian Inference in Spatiotemporal Cox Processes Driven by Multivariate Gaussian Processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (1): 157–75.

Gotway, Carol A, and Linda J Young. 2007. "A Geostatistical Approach to Linking Geographically Aggregated Data from Different Sources." *Journal of Computational and Graphical Statistics* 16 (1): 115–35.

References II

Grünbaum, Branko, and Geoffrey C Shephard. 1969. "Convex Polytopes." *Bulletin of the London Mathematical Society* 1 (3): 257–300.

Lee, Duncan. 2013. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software* 55 (13): 1–24.

Leroux, Brian G, Xingye Lei, and Norman Breslow. 2000. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 179–91. Springer.

Moraga, Paula, Susanna M Cramb, Kerrie L Mengersen, and Marcello Pagano. 2017. "A Geostatistical Model for Combined Analysis of Point-Level and Area-Level Data Using Inla and Spde." *Spatial Statistics* 21: 27–41.

Paumard, José. 1997. "Robust Comparison of Binary Images." *Pattern Recognition Letters* 18 (10): 1057–63.

Rue, Havard, and Leonhard Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Chapman; Hall/CRC.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–92.

Shephard, GC, and RJ Webster. 1965. "Metrics for Sets of Convex Bodies." *Mathematika* 12 (1): 73–88.

References III

White, Philip, Alan Gelfand, and Theresa Utlaut. 2017. "Prediction and Model Comparison for Areal Unit Data." *Spatial Statistics* 22: 89–106. https://doi.org/https://doi.org/10.1016/j.spasta.2017.09.002.