# BAYESIAN NONPARAMETRIC ANALYSIS OF SPATIAL VARIATION WITH DISCONTINUITIES

Cecilia Balocchi

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Co-Supervisor of Dissertation

Edward I. George, Professor of Statistics

Co-Supervisor of Dissertation

Shane T. Jensen, Professor of Statistics

Graduate Group Chairperson

Nancy R. Zhang, Professor of Statistics

Dissertation Committee

Bhaswar B. Bhattacharya, Assistant Professor of Statistics

James E. Johndrow, Assistant Professor of Statistics

BAYESIAN NONPARAMETRIC ANALYSIS OF SPATIAL VARIATION WITH DISCONTINUITIES

*Dedicated to my grandmothers, nonna Dirce and nonna Giulietta*

# ACKNOWLEDGEMENT

This thesis could not have been made without the constant support and encouragement of my advisors, Ed and Shane. Your intuition and wisdom have guided me in these five years. Your patience and your dedication has been invaluable. And you have been the best "cheerleaders" I could have ever hoped for, giving me positivity and confidence when I needed it most. Thank you.

I also would like to thank my committee members, Bhaswar and James, for the insightful conversation and the open doors. It has been a pleasure and honor working with you.

My thanks go to the entire Wharton Statistics Department, for fostering a friendly and welcoming environment. And a special 'thank you!' the wonderful members of the staff: your help has been priceless but most importantly you have been like a family away from home.

I am grateful to the wonderful friends and colleagues I have met during my time at Penn. Gemma and Sameer, Mo and Matteo, Daniela, Kathy and Federica, I feel so lucky to have shared this journey with you.

A big thanks to my parents, who have supported me in this adventure, even if it led me on the other side of the ocean. And finally thanks to Alberto, who has been present from day one, has always supported and encouraged me and has made these years so much more enjoyable.

# ABSTRACT

## BAYESIAN NONPARAMETRIC ANALYSIS OF SPATIAL VARIATION WITH DISCONTINUITIES

Cecilia Balocchi

Edward I. George, Shane T. Jensen

Spatial data often display high levels of smoothness but can simultaneously present abrupt discontinuities, especially in urban environments. In this dissertation we adopt a Bayesian perspective to account for these two contrasting facts, using partitions of areal data, and we then focus on three challenges that arise in this setting. First, we consider the applied problem of modeling crime trends over time in Philadelphia, measured at a local neighborhood level. We find that spatially local shrinkage imposed by a conditional autoregressive (CAR) model has substantial benefits in terms of out-of-sample predictive accuracy of crime. We also detect spatial discontinuities between neighborhoods that represent barriers. Then, we extend our search for barriers by clustering areal data. We propose a model that induces smoothness within clusters but allows for discontinuities between them, by assuming a "CAR-within-clusters" structure. The first challenge introduced by spatial clustering is that the combinatorially vast space of partitions makes typical stochastic search techniques computationally prohibitive. We introduce an ensemble optimization procedure that summarizes the posterior by simultaneously targeting several high probability partitions. We show on simulated data that our method achieves good estimation and partition selection performance. On the Philadelphia data we find that many recovered borders coincide with natural or built man-made barriers. The second challenge consists in choosing a distribution over partitions: standard distributions for exchangeable partitions are not appropriate for spatial data. We review and compare the properties of distributions for partitions of areal data that have been proposed in the literature and introduce new ones that display favorable properties. The third challenge relates to the problem of working with multiple

granularities: fixing one resolution can be restrictive because different granularities can be appropriate for different parts of a city. We introduce a model that combines the Nested Dirichlet Process with the Hierarchical Dirichlet Process to allow for flexible partitions of multi-resolution data and sharing of information between the partitions at different resolutions. We demonstrate our method on synthetic data and on real data in West Philadelphia, where central and suburban areas seem to be better represented by higher and lower resolutions, respectively.

TABLE OF CONTENTS

# Chapter 1

# Introduction

The spatial distribution of crime in urban environments suggests the validity of Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things". However, at the same time, it is evidence of a less popular but not less important statement by the geographer: "There is also obvious evidence that one must be carefully critical in applying the first law of geography. Anisotropic effects do occur, and so do discontinuities" (Tobler, 2004).

In this thesis, we focus on these two seemingly contradictory facts. We study how to combine the smoothness that characterizes spatial data with the discontinuities that are often present in urban environments.

Motivated by the modeling of crime in the City of Philadelphia, we first study the spatial distribution of crime trends over time. Accurate modeling of urban crime dynamics can offer various benefits: law enforcement officials can use this information to decide how to deploy resources to ensure public safety, urban planners can gain knowledge on crime is affected by socio-economic factors and the built environment, and city officials can improve the quality of life in the city with community programs and interventions. We consider yearly counts of violent crime data in the years of 2006-2015, aggregated at a local neighborhood resolution. We model the spatio-temporal crime dynamics using a linear model that allows neighborhood-specific behaviors. We study the spatial distribution of the neighborhood-specific mean level of crime and trends over time.

Data that is measured within fixed spatial regions is called *areal data*, with the regions often referred to as *areal units*. This is different from *point-referenced* data, which instead are defined by coordinates that can vary continuously over the space, such as GPS-coordinates. While the spatial aspect of point-referenced data can be described simply by their location in

space, the geographic structure of areal data is characterized by the adjacency relationship between the areal units. Two regions can share a border and thus be neighbors, or not.

Using models for areal data, we explore how sharing information between neighboring regions improves estimation. We then extend these spatially smoothing models to allow for discontinuities: we aim at detecting *barriers*, i.e. borders between neighborhoods that identify discontinuities. We consider two different approaches: we first recover barriers as local segments separating two neighborhoods; we then identify them as closed lines that separate whole regions from the surrounding areas, by partitioning the neighborhoods into *clusters* of contiguous units displaying similar behaviors.

In our attempt to model spatial variation and recover partitions, we take a Bayesian perspective, which provides us with various benefits. When modeling spatial smoothness, it allows us to easily incorporate "sharing of information" between areal units with hierarchical models. In the search for partitions of areal units, we can introduce prior information and account for their uncertainty encoded in the posterior distribution, by considering the parameters as random variables.

In this thesis, we consider the applied problems of modeling and clustering crime trends in Philadelphia and we tackle several challenges that clustering areal data poses.

In chapter 2, we consider crime trends measured within local neighborhoods, and compare models with no shrinkage, global shrinkage and spatial local shrinkage, the latter defined using conditionally auto-regressive (CAR) models Besag (1974). We show that local shrinkage models achieve lowest out-of-sample prediction errors. We also explore the first local approach to recovering barriers, following the tradition of *wombling* (Womble, 1951). We propose a model that identifies them as borders between pairs of neighborhoods between which we should not be sharing informations, while accounting for uncertainty. This chapter has been adapted from the research article "Spatial modeling of trends in crime over time in Philadelphia" (Balocchi and Jensen, 2019).

In chapter 3, we consider the second approach of finding barriers; this approach identifies them as the closed-lines borders between clusters of neighborhoods. We propose a model that partitions the city and induces smoothness within clusters but allows for discontinuities between them, by assuming a "CAR-within-clusters" structure. In terms of model implementation, conventional stochastic search techniques are computationally prohibitive, as they must traverse a combinatorially vast space of partitions. Rather than directly sampling from the posterior distribution of the discrete partitions, we summarize the posterior identifying several partitions with largest posterior probability. We achieve this by extending the ensemble optimization procedure introduced in Ročková (2018). At a very high level, this procedure runs several greedy searches over the discrete posterior distribution that are made "mutually aware" through an entropy penalty that promotes diversity among the search trajectories. On simulated and real data, our proposed method demonstrates good estimation and partition selection performance. We use the identified partitions to estimate crime trends in Philadelphia. This chapter has been adapted from the work "Bayesian Spatial Clustering of Crime in Philadelphia with Particle Optimization" with Sameer Deshpande, Ed George and Shane Jensen (Balocchi et al., 2019).

In chapter 4, we consider the problem of specifying a prior distribution for partitions of areal data. A common choice in the Bayesian nonparametric literature is the distribution induced by the Dirichlet Process, because of its mathematical convenience; however, the property that makes it so attractive, the fact of inducing exchangeable partitions, is also the one that makes inappropriate for describing spatial data: spatial data is not exchangeable and the probability that two adjacent units belong to the same cluster should not be the same as that for two units that are distant. In this chapter we review and compare the properties of distributions for partitions of areal data that have been proposed in the literature, and introduce new ones that follow the framework of Müller et al. (2011).

Finally, in chapter 5, we consider the problem of multi-resolution clustering. Cities like Philadelphia, and more generally regions divided into areal units, can be split at different

granularities. The majority of the analyses of areal data fix one layer of resolution and only study the chosen granularity. However, in complex environments, such as cities, the spatial variation of the data can arise at different resolutions in different regions. For example, central areas are more densely populated and heterogeneous, compared to suburban areas. In central areas we might expect a phenomenon like crime to be changing at a finer granularity than in suburban areas. Therefore, instead of reducing the analysis to a specific level, it can be beneficial to simultaneously consider multiple resolutions. We introduce a new model, the nested Hierarchical Dirichlet Process, that allows clustering of areal units at multiple resolutions, while sharing information between the different levels partitions.

The rest of this introductory chapter reviews basic concepts of spatial data analysis and nonparametric Bayesian statistics.

## 1.1. Spatial Data Analysis Introduction

Spatial data can be classified into three types:

- *point-referenced* data, for which the location varies continuously over the space;
- *areal* data, for which the location varies over a discrete and finite set of well-defined regions that partition the space;
- *point-process* data, for which the location is itself random, i.e. the support of a spatial point-process.

A common method for modeling spatial point-referenced and point-processes data is kriging or Gaussian process interpolation (Stein, 2012; Cressie, 1990). This can be studied either with a classical approach, or with a Bayesian approach (Banerjee et al., 2014). Additional popular models for point-processes data consider other frameworks such as Gibbs point processes, Poisson processes and Cox processes; see Møller and Waagepetersen (2007).

Common classical methods for modeling areal data are spatial autoregressive models, that

include the Simultaneous Autoregressive Model (Whittle, 1954), the spatial Durbin model (Anselin, 2013) and the Conditionally Auto Regressive model (Besag, 1974); for a review of these and other methods see LeSage and Pace (2009). Many of these models have also been considered and used in a Bayesian framework (Banerjee et al., 2014).

### 1.1.1. Areal data

Spatial areal data is measured over a discrete and finite set of regions that partition the space. These regions, known as *areal units*, have well-defined and fixed borders, which determine the adjacency relationship between the units.

The geographical structure of the areal units is encoded in the matrix W= $(w_{ij})$ of weights that capture the spatial proximity of the areal regions. Often, the proximity is defined as the adjacency relationship and the matrix W corresponds to the *adjacency matrix*:

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share a border,} \\ 0 & \text{if } i \text{ and } j \text{ do not share a border.} \end{cases}$$

"Sharing a border" can mean different things, depending on the contiguity method used: according to the *rook* contiguity method, two regions are adjacent if their borders share at least a segment; according to the *queen* method instead, two regions share a border if they share at least a point.

To measure the strength of this spatial correlation of areal data, one of the standard statistics is Moran's $I$ (Moran, 1950; Banerjee et al., 2014), which is defined as

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

Moran's $I$ can be used for testing for spatial autocorrelation: under the null hypothesis of no spatial association, it is possible to compute exactly the mean (equal to $-\frac{1}{n-1}$) and standard error of Moran's $I$ and construct a Z-test.

### 1.1.2. Conditionally Auto Regressive models

To model areal data or random variables we will mostly use the conditionally autoregressive (CAR) model, which was introduced in its most general formulation by Besag (1974). The CAR model is a Gaussian Markov random field which induces spatial dependence through an adjacency matrix for the areal units.

Several variations of this CAR framework are reviewed and compared in Lee (2011). In this work, we will use the proper CAR formulation introduced by Leroux et al. (2000).

Let $\boldsymbol{\theta}$ denote a vector of elements that are potentially spatially correlated, where each component $\theta_i$ corresponds to the parameter for areal unit $i$. Leroux et al. (2000) defines the distribution of each $\theta_i$ given the other $\boldsymbol{\theta}_{-i}$ as a normal distribution centered at a weighted average of a global mean and the $\theta_j$'s from bordering neighborhoods,

$$\theta_i \mid \boldsymbol{\theta}_{-i}, \theta_0, \tau^2 \sim \mathrm{N}\left(\frac{\rho \sum_j w_{ij}\,\theta_j + (1-\rho)\,\theta_0}{\rho \sum_j w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho \sum_j w_{ij} + (1-\rho)}\right),$$

where $w_{ij}$ are adjacency weights that are equal to 1 if the neighborhoods $i$ and $j$ share a border and equal to 0 otherwise.

The parameter $\rho \in [0,1]$ represents the strength of the spatial correlation between the components of $\boldsymbol{\theta}$, where larger values of $\rho$ correspond to a stronger influence of bordering neighborhoods. In the special case of $\rho = 0$, the CAR distribution reduces to the multivariate normal distribution with independent covariance structure.

It can be proved (Banerjee et al., 2014, Ch.3) using Brook's lemma (Brook, 1964), that the joint distribution of $\boldsymbol{\theta}$ is uniquely determined by the set of conditional distributions:

$$\boldsymbol{\theta}|\theta_0, \tau^2 \sim \mathrm{N}\left(\theta_0 \cdot \mathbf{1}\,,\, \tau^2 \cdot [\rho(\mathbf{D}_W - \mathbf{W}) + (1-\rho)\mathbf{I}]^{-1}\right)$$

where $\mathbf{1}$ is a vector of 1's and $\mathbf{D}_W - \mathbf{W}$ is the Laplacian matrix based on our neighborhood

adjacency matrix $\mathbf{W}$. For values of $\rho$ in $[0, 1)$ the joint distribution is proper, while for $\rho = 1$ the distribution is degenerate (Lee, 2011). By adding the constraint $\sum_i(\theta_i - \theta_0) = 0$ we can get a distribution for a $n$-dimensional vector, concentrated in a $(n-1)$-dimensional subspace; this is known as the intrinsic CAR by Besag et al. (1991).

## 1.2. Bayesian Nonparametrics Introduction

Clustering is often performed with a model-based approach that relies on *mixture modeling*. In parametric Bayesian mixture models, each data point is generated from one of a finite set of mixtures; the mapping of data into mixtures identifies a partition of the units into clusters. However, these models require specification of the number of mixtures *a priori*, which is often unknown. Bayesian nonparametric mixture models instead do not assume a fixed, finite number of mixtures, but rather an unbounded number of mixtures, which allows for the number of clusters in the data to be adaptively estimated.

Many nonparametric Bayesian mixture models are based on the Dirichlet Process (Ferguson, 1973), which is a distribution over random probability measures. Let $G_0$ be a distribution on $(\Theta, \mathcal{B})$ and $\alpha > 0$. We say that $G$, a random probability measure on $(\Theta, \mathcal{B})$, is distributed according to the Dirichlet Process $\mathrm{DP}(\alpha, G_0)$, if for any finite partitions $(A_1, \ldots, A_k)$ of $\Theta$,

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}\left(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)\right).$$

In this case we write $G \sim \mathrm{DP}(\alpha, G_0)$. The parameters that characterize the DP, $\alpha$ and $G_0$, are respectively called the concentration parameter and the base distribution. Ferguson (1973) showed that a draw $G$ for the Dirichlet Process is almost surely discrete and can be written as

$$G = \sum_{k=1}^{\infty} p_k \delta_{\theta_k^*},$$

where $\delta_{\theta_k^*}$ are the atoms of $G$ and $p_k$ is the probability associated with $\theta_k^*$. According to the *stick-breaking construction* of $G$ (Sethuraman, 1994), the atom locations $\theta_k^*$ are i.i.d.

random variables distributed according to $G_0$ and the probabilities $p_k = b_k \prod_{j=1}^{k-1}(1-b_j)$, where $b_j \overset{iid}{\sim} \text{Beta}(1, \alpha)$; we will denote the stick-breaking construction with $(p_k) \sim SB(\alpha)$.

Because of its discreteness, the Dirichlet Process is rarely employed to directly model data; instead, it is mostly used for specifying a prior distribution for the mixture components in mixture modeling. Let $\{y_1, \ldots, y_n\}$ be a set of exchangeable observations drawn from a mixture model, and let $\{\theta_1, \ldots, \theta_n\}$ be the latent mixture components associated with each observation: $y_i$ given $\theta_i$ is drawn from $f(\cdot; \theta_i)$. To specify a prior distribution on the latent mixture components we assume $\theta_1, \ldots, \theta_n | G \overset{iid}{\sim} G$ and $G \sim DP(\alpha, G_0)$.

$$
\begin{aligned}
y_i | \theta_i &\sim f(\cdot; \theta_i) \\
\theta_i | G &\sim G \\
G | \alpha, G_0 &\sim \text{DP}(\alpha, G_0).
\end{aligned}
$$

For the rest of this description we will mostly focus on the modeling of the latent mixture components, and we will assume that the observations are drawn from a parametric distribution $f$ parametrized by the mixture component.

The discreteness of $G$ also implies that among the latent mixture components $\theta_1, \ldots, \theta_n$ some values will be repeated with high probability. Therefore, a partition $\gamma$ is induced on the data by the mixture components, with clusters identified by their unique values $\theta^*_{i_1}, \ldots, \theta^*_{i_K}$: for $k = 1, \ldots, K$, we define $S_k = \{i : \theta_i = \theta^*_{i_k}\}$ and $\gamma = \{S_1, \ldots, S_K\}$.

Before analyzing the behavior of the mixture components, let us describe the posterior distribution of $G$ given $\theta_1, \ldots, \theta_n$. Let $(A_1, \ldots, A_k)$ be a partition of $\Theta$ and let $n_k = |\{\theta_i \in A_k\}|$. It is easy to see that

$$
(G(A_1), \ldots, G(A_k)) | \theta_1, \ldots, \theta_n \sim \text{Dir}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_k) + n_k).
$$

Then the posterior distribution of $G$ is a Dirichlet Process, with updated parameters:

$$G|\theta_1, \ldots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right).$$

If now we consider the conditional predictive distribution of $\theta_i|\theta_1, \ldots, \theta_{i-1}$, marginally of $G$, we can recover the Blackwell-MacQueen Urn (Blackwell et al., 1973):

$$\begin{aligned}\Pr(\theta_i \in A|\theta_1, \ldots, \theta_{i-1}) &= \mathbb{E}\left[G(A)|\theta_1, \ldots, \theta_{i-1}\right] = \\ &= \frac{1}{\alpha + n}\left(\alpha G_0(A) + \sum_{j=1}^{i-1} \delta_{\theta_i}(A)\right)\end{aligned}$$

It is easy to see that the sequence $\theta_1, \ldots, \theta_n$ is exchangeable. As a consequence, also the partition $\gamma$ is exchangeable, i.e. its probability is invariant to permutation of the units indexes. Moreover, its probability $p(\gamma = \{S_1, \ldots, S_K\})$ can be described by the *exchangeable partition probability function* (EPPF) $p(n_1, \ldots, n_K)$, where $n_i$ is the size of cluster $S_i$:

$$p(\gamma = \{S_1, \ldots, S_K\}) = p(n_1, \ldots, n_K) = \frac{\alpha^k \prod_{j=1}^{k}(n_j - 1)!}{\alpha(\alpha + 1)\ldots(\alpha + n - 1)}$$

which is also known as Ewens' formula (Ewens, 1972; Pitman, 1995).

The same distribution over random partitions can be achieved with the *Chinese Restaurant Process* (CRP) (Aldous, 1985) iterative construction. In the CRP metaphor, each unit is represented by a costumer entering a restaurant with an infinite number of tables, who picks at which table to seat. Each table corresponds to a cluster and each costumer chooses proportionally to the number of people already sitting at the table, or chooses a new table with probability proportional to $\alpha$. Let $z_i$ represent the cluster membership of unit $i$ (or its table choice), let $K$ be the number of clusters occupied by the first $i - 1$ units and $n_k$

be the size of cluster $k$: $n_k = |\{i : z_i = k\}|$. Then,

$$p(z_1 = 1) = 1$$
$$p(z_i | z_1, \ldots, z_{i-1}) = \sum_{k=1}^{K} \frac{n_k}{i - 1 + \alpha} \delta_k + \frac{\alpha}{i - 1 + \alpha} \delta_{K+1}$$

This distribution is characterized by a *rich-get-richer* behavior, as a large cluster attracts even more units. As a consequence, as the number of units $n$ increases, the number of clusters grows at a logarithmic rate, as $\alpha \log(n)$.

The mathematical tractability of Ewens' formula, of the Blackwell-MacQueen Urn and of the Chinese Restaurant Process, have made the Dirichlet Process widely used. However, the rich-get-richer behavior that the Dirichlet Process induces might not be suitable for some situations. The Pitman-Yor process (PYP) (Pitman, 1995; Pitman and Yor, 1997) is another distribution on random probability distributions. The distribution it induces over partitions allows for a larger number of clusters, and an asymptotic power-law growth. This distribution, also known as the Ewens-Pitman distribution, is described by the Ewens-Pitman formula:

$$\Pr(\gamma = \{S_1, \ldots, S_k\}) = \frac{\prod_{j=1}^{k-1} (\alpha + j\sigma)}{(\alpha + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}$$

where $\sigma \in [0, 1)$ and $\alpha > -\sigma$. This distribution includes the special case of the Ewens' formula, when $\sigma = 0$.

The Ewens-Pitman distribution can also be recovered from the sequence of conditional distributions that extend the Chinese Restaurant Process. If $z_i$ represent the cluster membership of unit $i$, $K$ is the number of clusters occupied by the first $i - 1$ units and $n_k$ the

size of cluster $k$, then

$$p(z_1 = 1) = 1$$
$$p(z_i | z_1, \ldots, z_{i-1}) = \sum_{k=1}^{K} \frac{n_k - \sigma}{i - 1 + \alpha} \delta_k + \frac{\alpha + K\sigma}{i - 1 + \alpha} \delta_{K+1}$$

The new parameter $\sigma$ is affecting the number of clusters: the asymptotic number of clusters grows as $n^\sigma$. This *power law* behavior is sometimes considered more appropriate for certain real world applications.

# Chapter 2

# Spatial Modeling of Trends in Crime over Time

## 2.1. Introduction

Modeling and prediction of crime has always been of interest to local authorities, police departments and governments to assure safety of the population and more efficient law enforcement. Recent availability of detailed crime data has made this effort even more accessible to statistical practitioners and the general public.

As an example, the Philadelphia police department has released detailed information about reported crimes committed from 2006 to the present day[1]. The information about each reported crime includes the type of crime (which we will describe in Section 2.2), the date and time of the crime and the GPS location of the crime.

Using their reported crime data, many police departments have used statistical modeling procedures and algorithms to help predict locations of crimes for better prevention and faster intervention (Hvistendahl, 2016). The modeling of crime locations is not only useful for law enforcement but also for marketing strategies related both to real estate and commercial activities, e.g. Trulia[2] uses crime data as part of their evaluation of the relative safety and attractiveness of different neighborhoods. In this work, we will focus on estimating changes in violent crimes over the past decade at a local neighborhood resolution which will involve both temporal and spatial modeling of crime.

Many different approaches have been taken to the modeling of the spatial distribution of crime. These approaches can be subdivided into two general categories, either modeling

Adapted from a research article:
**Balocchi, C.**, Jensen, S. T. (2019) "Spatial modeling of trends in crime over time in Philadelphia" *Annals of Applied Statistics* 13(4):22352259

[1] http://www.phlcrimemapper.com/

[2] https://www.trulia.com/

crime as a spatial point process using the specific locations of each reported crime (Mohler et al., 2011; Taddy, 2010; Flaxman, 2014) or modeling crime as *areal* data, i.e. totals aggregated within larger regions, as in Aldor-Noiman et al. (2016), Law et al. (2014) and Li et al. (2014).

A common method for modeling spatial point processes is kriging or Gaussian process interpolation (Stein, 2012; Cressie, 1990). This can be studied either with a classical approach, or with a Bayesian approach (Banerjee et al., 2014). Alternative popular models consider other frameworks such as Gibbs point processes, Poisson processes and Cox processes; see Møller and Waagepetersen (2007).

Common classical methods for modeling areal data are spatial autoregressive models, that include the Simultaneous Autoregressive Model (Whittle, 1954), the spatial Durbin model (Anselin, 2013) and the Conditionally Auto Regressive model (Besag, 1974); for a review of these and other methods see LeSage and Pace (2009). Many of these models have also been considered and used in a Bayesian framework (Banerjee et al., 2014).

Our goal in this work is the estimation of trends in violent crime over the past decade at a high resolution local neighborhood level throughout the city of Philadelphia. As it is well established that crime frequencies are spatially correlated (Herbert, 1982; Brantingham and Brantingham, 1984), we need to create a model that allows the change in crime over time to be correlated by locally proximal neighborhoods. Our model will also account for characteristics of each local neighborhood, including the population count of the area and economic health of residents, as measured by median income and poverty level of households.

In addition to aiding law enforcement, accurate estimation of changes in crime at the local neighborhood level would also enable the study of the association between crime trends and changes in the built environment. We are particularly interested in how aspects of the built environment encourage *vibrancy*, a measure of positive human activity, and how vibrancy is associated with safety in local neighborhoods (Humphrey et al., 2017).

The city of Philadelphia is a particularly interesting case study for estimating trends in crime as it is a large urban area that is currently undergoing substantial development and experiencing population growth for the first time in decades. In addition to our primary goal of estimation of changes in crime in Philadelphia neighborhoods, this application also provides an interesting spatio-temporal data context for comparing different Bayesian shrinkage approaches to spatial areal modeling.

We will take an areal approach to modeling crime since our primary goal is greater understanding of evolving crime dynamics at the local neighborhood level within the city of Philadelphia. Our areal units will be U.S. Census block groups which consist of 10-20 city blocks and which are naturally interpretable as neighborhoods. U.S. Census block groups are also the highest resolution for which economic data is available as covariate information.

Compared to previous areal approaches (e.g. Aldor-Noiman et al. (2016), Law et al. (2014) and Li et al. (2014)), we are using smaller areal units and we will focus on not only total crime but also the trend in crime over time within each local neighborhood. We have a longer time period (ten years) of recorded crimes for estimating time trends than Law et al. (2014) that worked with property crimes over a two year period.

Our methodological contribution is the development of a Bayesian spatial modeling framework to explore global vs. local smoothing for our parameter estimates while also allowing for data-driven discontinuities in our model between proximal areal units. Using a Bayesian approach allows us to induce this smoothing through shrinkage priors for our parameters and also enables us to estimate borders between neighborhoods that have a high probability of being barriers.

In Section 2.2, we provide details for the neighborhood structure of Philadelphia and describe the detailed crime data that we will use to estimate changes in crime over the past decade. We also outline the demographic, economic and land use measures we will use as neighborhood-level predictors of violent crime in our spatial models. The code for acquiring

and cleaning the data that were used in this analysis is available as a GitHub repository at `https://github.com/cecilia-balocchi/Urban-project`. In Section 2.3, we develop several Bayesian modeling approaches for global or local sharing of information between Philadelphia neighborhoods, as well as a model extension that allows for spatial discontinuities in our parameter estimates between proximal neighborhoods. We then compare these modeling options in terms of both in-sample and out-of-sample predictive accuracy in Section 2.4. We visualize and discuss the results of our spatial modeling of crime trends for Philadelphia in Section 2.5 and then conclude with a brief discussion in Section 2.6.

## 2.2. Population, Economic and Crime Data in Philadelphia

The population and economic data are provided by the US Census Bureau whereas crime data is provided by the Philadelphia Police Department. Our definition of local neighborhoods in Philadelphia will be based upon the "block group" geographical units defined by the US Census Bureau. The city of Philadelphia is divided into 384 census tracts which are divided into 1336 block groups. Shapefiles from the US Census Bureau give the boundaries and area of each census block group. Figure S1 in appendix A gives a map outlining the 1336 block groups in Philadelphia.

Our motivation for analyzing trends in crime at this resolution is two-fold: a. US census block groups consist of 10-20 city blocks which generally matches our concept of a "neighborhood" and b. the block group level is the highest resolution of the economic data that we will use as predictors of crime. The average size of block groups in Philadelphia is 0.26 $km^2$, with an average population of 1142 residents.

Our population data was pulled from the census website[3] by setting the geography as all blocks in Philadelphia and setting the data source as "Hispanic or Latino Origin By Race" (which is SF1 P5 in their database). The raw demographic data gives the population count in each block group from the 2010 census. Figure S1 in appendix A gives the population

[3]https://factfinder.census.gov/

count for each block group in Philadelphia.

The same data also has the population count in each block group divided by ethnic categories[4]. From these ethnicity counts, we calculate a measure of the *segregation* in each block group as

$$\text{segregation}_i = \tfrac{1}{2} \sum_r |p_{i,r} - \overline{p}_r|$$

where $p_{i,r}$ is the proportion of ethnicity $r$ in block group $i$ and $p_r$ is the proportion of ethnicity $r$ across the entire city of Philadelphia. The fraction $\frac{1}{2}$ scales this segregation measure to be between 0 and 1.

In addition to population count and our segregation measure, we will also consider several measures of the economic health of each neighborhood. Our economic data comes from the American Community Survey from the same US census website as our population data, specifically tables B19301 for income and C17002 for poverty, both from 2013. This data is only available at the resolution of census block groups. For each block group (neighborhood) in Philadelphia, we have income per capita as one predictor of crime.

We also have information about the proportion of households in various states of poverty. Specifically, we have the fraction of the population in seven different brackets of income-to-poverty-line ratios: $[0, 0.5)$, $[0.5, 1)$, $[1, 1.25)$, $[1.25, 1.5)$, $[1.5, 1.85)$, $[1.85, 2)$, $[2, \infty)$. For example, the $[0.5, 1)$ bracket represents families that have income between 50% of the poverty line and the poverty line itself. The poverty line is defined by the Census Bureau according to the size and composition of a household (e.g. a family with two children has a poverty line threshold of $23,999).

We use this poverty data to create a single measure of poverty for each block group (neighborhood) by calculating a weighted sum of the proportion of households in each of the seven

[4]The ethnic categories are: White, Black, Asian, Native Americans, Native Pacific Islanders (including Hawaii), Other, Two or more races (nonhispanic) and Hispanic/Latino. We combined Native Americans, Native Pacific Islanders, and Two or more races into the Other category, which leads to five ethnicities in our analysis: 1. White, 2. Black, 3. Hispanic, 4. Asian, and 5. Other.

poverty brackets:

$$\text{poverty}_i = \sum_{j=1}^{7} w_j \, q_{i,j}$$

where $q_{i,1}$ is the proportion of households in block group $i$ that are in the lowest bracket $[0, 0.5)$ and $q_{i,7}$ is the proportion of households in block group $i$ in the highest bracket $[2, \infty)$. We use linearly decreasing weights $\mathbf{w} = [1, 5/6, 4/6, 3/6, 2/6, 1/6, 0]$ to give higher weight to the brackets with higher poverty. Our poverty measure varies from 0 to 1, with larger values implying higher poverty.

In addition to the demographic and economic predictors described above, we also derive measures of the *built environment* that may also be predictive of crime. Our data on the built environment comes from the zoning designation of each lot in Philadelphia. Zoning data from the City of Philadelphia provides the area and registered land use designation (e.g. commercial, residential, industrial, vacant, transportation, park, civic) of all 560,000 lots in Philadelphia.

We create several land use metrics from these zoning designations that could be predictive of crime. First, we calculate the fraction of area in each block group $i$ that is designated as 'Vacant',

$$\text{vacancy}_i = \frac{Area_i(\text{Vacant})}{Area_i}$$

Second, we calculate the ratio of the area in each block group $i$ that is commercial versus residential,

$$\text{comresprop}_i = \frac{Area_i(\text{Commercial})}{Area_i(\text{Commercial}) + Area_i(\text{Residential})}$$

To summarize, we have created six neighborhood characteristics that we will use as predictors of crime: population count, segregation, median household income, poverty, vacant proportion and commercial vs. residential proportion. Some block groups in Philadelphia have missing values for the economic predictors due to a very small or zero population count. We exclude these block groups (a total of eight) from our analysis. We additionally exclude one block group containing the detention centers in Philadelphia.

Our crime data comes from the Philadelphia Police Department and includes all crimes reported by the police in the city of Philadelphia from January 1, 2006 to December 31, 2015. For each reported crime, we have the type of crime, the date and time of the crime, and the location of the crime in terms of the GPS latitude and longitude (WGS84 decimal degrees). Each crime in our dataset is categorized into one of several types: homicide, sex crime, armed robbery, assault, burglary, theft, motor vehicle theft, etc.

We make a distinction between *violent* and *non-violent* (property) crimes in our analysis. As defined by the Uniform Crime Reporting program of the FBI, *violent* crimes include homicides, rapes, robberies and aggravated assaults whereas *non-violent* crimes include burglaries, thefts and motor vehicle thefts.

Our own crime categorization differs from the FBI in two ways. We combine 'rapes' and 'sex assaults' (which changed in definition in 2013) into a broader 'sex crimes' category and consider all 'sex crimes' as violent crimes. The FBI also makes a distinction between 'aggravated assaults' and 'other assaults', with the latter being where an injury does not occur but the threat of injury is present. In contrast, we combine both 'aggravated assaults' and 'other assaults' into a broader 'assaults' category and consider all 'assaults' as violent crimes.

For this work, we focus entirely on the modeling of violent crimes as they have the most direct impact on human safety and the perception of safety. However, non-violent crimes are also important to track for law enforcement and are a focus of ongoing research. In the subsequent analyses in this work, we will use 'crime' to mean only violent crimes.

In Figure 1, we give the counts of each type of violent crime within each year in 2006-2015, aggregated over the entire city. We see generally decreasing trends within the assault and robbery categories, which are the most numerous types of crimes. Sex crimes and homicides are also somewhat decreasing over this time span though it is harder to see this trend given the low counts for either type of crime.

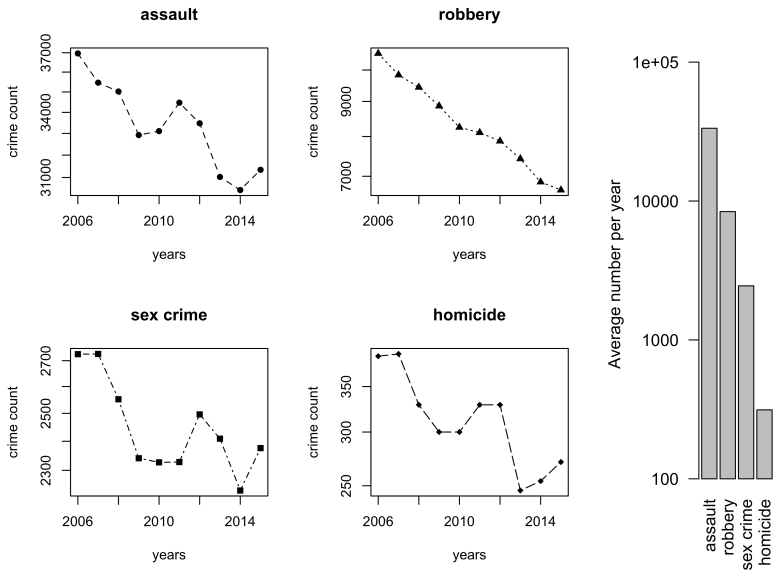


Figure 1: Counts of the different types of violent crimes in each year aggregated over the entire city of Philadelphia.

Clearly, the impression given from Figure 1 is that violent crimes are generally decreasing in the city of Philadelphia over the time period from 2006 to 2015. However, are there specific neighborhoods that show substantially larger decreases or even some neighborhoods that show increases in violent crimes in this period?

As discussed in Section 2.1, we will model the spatial distribution of crime with an areal approach where our areal units are U.S. Census block groups which we define as the local neighborhoods of Philadelphia. Violent crimes are aggregated within each U.S. Census block group based on the GPS coordinates of each reported crime.

One issue with this approach is that some crimes occurring near to a boundary between U.S. Census block groups could be aggregated into the incorrect areal unit due to measurement error or ambiguity in their recorded point locations. This possibility is one of several motivations for our hierarchical Bayesian modeling approach that shares information between adjacent block groups when estimating crime totals and trends in crime over time across the city of Philadelphia.

In Figure S2 of appendix A, we give the count of violent crimes per year in each block group averaged over the years 2006-2015. One can see substantial heterogeneity across block groups in the average counts of violent crimes per year. There are several outlying values: particular block groups that have much higher average violent crime counts.

These outlying neighborhoods motivate us to examine violent crime totals on the log scale. In Figure S2 of appendix A, we also give the average of the logarithm of the count of violent crimes per year in each block group, averaged over the years 2006-2015. We can see more details of the spatial distribution of violent crime on the log scale. Modeling crime on the log scale has the additional benefit that changes in log crime can be interpreted as percentage changes in crime.

We also see in Figure S2 evidence of spatial correlation in violent crime totals between proximal block groups throughout the city. This is not surprising since the factors that lead to crime likely vary throughout the city in a (mostly) spatially continuous fashion. It is this spatial correlation that will be the focus of our modeling work in Section 2.3.

To get an idea of the strength of this spatial correlation, one of the standard statistics used for areal data is Moran's $I$ (Moran, 1950; Banerjee et al., 2014), which is defined as

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where $W = (w_{ij})$ is a matrix of weights that capture the spatial proximity of the areal regions. We set $w_{ij}$ to be 1 if block groups $i$ and $j$ share a border and 0 otherwise. We use the *queen* contiguity method so two block groups share a border if they share at least a point on their boundaries.

Moran's $I$ can be used for testing for spatial autocorrelation: under the null hypothesis of no spatial association, we can compute exactly the mean (equal to $-\frac{1}{n-1}$) and standard error of Moran's $I$. Calculating $I$ on the total number of violent crimes from 2006 to 2015 in our data gives an observed value of 0.335, compared to a null mean of 0.0007 and standard

error of 0.0127, which suggests a highly significant amount of spatial autocorrelation in violent crime totals.

In the next section, we develop several different Bayesian strategies for modeling violent crime over time and spatially between the areal neighborhoods of Philadelphia. We will fit our models on the violent crime data from 2006 to 2014, leaving data from 2015 for model comparison and evaluation.

## 2.3. Modeling Areal Crime Data over Space and Time

As described in Section 2.2, the areal units of our analysis are the 1336 US census block groups of Philadelphia (shown in Figure S1 in appendix A).

For the remainder of this work, we will use the terms "block group" and "neighborhood" interchangeably. The input data for our analysis is the number of violent crimes, $c_{it}$, reported in year $t$ within neighborhood $i$. Our temporal range is $t = 1, \ldots, T$ with $T = 10$, for the years 2006-2015 and our spatial range is $i = 1, \ldots, n$ with $n = 1336$, for all the block groups in Philadelphia.

As seen in the violent crime totals (averaged over time) in Figure S2 of appendix A, there are some substantial outlying neighborhoods with high violent crime totals relative to most of the city. These outliers (and general skewness in violent crime totals) motivates us to model violent crime totals on the logarithmic scale. This strategy has the additional benefit that linear changes over time in the logarithm of violent crime totals can be interpreted as percentage changes in raw violent crime totals.

However, because there are a small number of neighborhoods with zero crimes in some years, we need to consider a transformation that is defined at zero. Accordingly, we use the inverse hyperbolic sine transformation (Burbidge et al., 1988b) that is centered to give values approximately equal to the logarithmic transformation. Specifically, we calculate our

transformed violent crime totals as

$$y_{it} = \log(c_{it} + \sqrt{c_{it}^2 + 1}) - \log(2) \tag{2.1}$$

where $c_{it}$ is the total number of violent crimes reported in year $t$ within neighborhood $i$. A more common solution would be to add a small non-zero value to the counts, e.g. $\log(c_{it}+1)$. We prefer the inverse hyperbolic sine transformation as it is numerically equivalent to the log transformation for large counts but is a better approximation than the $\log(c_{it} + 1)$ transformation for small counts.

An alternative modeling strategy for count data does not apply a transformation but assumes a Poisson distribution for the counts (Law et al., 2014; Li et al., 2014; Anderson and Ryan, 2017). The Poisson model would not work since our data is over-dispersed, and the more flexible negative-binomial distribution does not model mean and variance as intuitively as a normal model. In addition, the normal model is conjugate for the prior distributions we will be considering which eases posterior estimation.

### 2.3.1. Accounting for Neighborhood Level Covariates

We use a standard linear regression approach to account for the neighborhood-level economic, demographic and land use predictors of crime. Our transformed violent crime totals $y_{it}$ are modeled as,

$$y_{it} = \alpha + \mathbf{z}_i^T \boldsymbol{\gamma} + e_{it}, \tag{2.2}$$

where $\mathbf{z}_i$ is the vector of predictor variables for neighborhood $i$ and $\boldsymbol{\gamma}$ is the vector of coefficients for those predictor variables, so $\mathbf{z}_i^T \boldsymbol{\gamma} = \sum_{d=1}^{6} \gamma_d z_{id}$.

As outlined in Section 2.2, we have $d = 6$ predictor variables of crime for each neighborhood: population count, segregation, median household income, poverty, vacant proportion and commercial vs. residential proportion. We used square root transformations of vacant proportion, commercial vs. residential proportion and poverty and a logarithmic transfor-

mation of income to give a more linear relationship with the outcome variable.

Although yearly demographic and economic data is available after 2013, we avoid extrapolating values of the predictors to earlier years by modeling each predictor variable as *static* over the ten year period spanned by our crime data. We examine the estimated partial effects $\boldsymbol{\gamma}$ of these economic, demographic and land use predictors in Section 2.5.1.

Although there is interest in the partial effects of our crime predictors, our primary interest lies in the temporal trends captured by $e_{it}$ and the spatial correlation in these trends. With these time trends, we will be able to answer questions such as 'what areas of the city are increasing or decreasing most quickly in terms of safety?'.

### 2.3.2. Time Trends with No Spatial Correlation

We can add a global linear trend over time into our model,

$$y_{it} = \alpha + \mathbf{z}_i^T \boldsymbol{\gamma} + \beta \cdot t + \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim \mathrm{N}(0, \sigma^2) \tag{2.3}$$

where the scalar coefficient $\beta$ can be interpreted as the global percentage change in violent crime over time across the entire city of Philadelphia and $t$ takes on integer values from 1 to 10 to represent the years 2006-2015.

However, this model with only a global $\alpha$ and $\beta$ does not allow for heterogeneity between different neighborhoods in the overall level of violent crime or trend in violent crime over time. We can account for this heterogeneity through neighborhood-specific intercepts $\alpha_i$ and slopes $\beta_i$, which give us the model

$$y_{it} = \alpha_i + \mathbf{z}_i^T \gamma + \beta_i \cdot t + \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim \mathrm{N}(0, \sigma^2). \tag{2.4}$$

However, model (2.4) is over-parameterized: in fact, the effect of our static covariates is completely explained by the neighborhood-specific intercepts $\alpha_i$, so the same fit can be

achieved by removing the covariates,

$$y_{it} = \alpha_i + \beta_i \cdot t + \epsilon_{it} \quad \text{where} \quad \epsilon_{it} \sim \mathrm{N}(0, \sigma^2) \tag{2.5}$$

Nonetheless, we can still estimate the partial effects of the covariates with an equivalent two-stage approach where we first fit $y_{it} = \alpha + \mathbf{z}_i^T \gamma + e_{it}$ and then fit the estimated residuals with the neighborhood-specific coefficient model, $\hat{e}_{it} = \alpha_i + \beta_i t + \epsilon_{it}$.

These neighborhood-specific model coefficients allow us to identify regions of Philadelphia with different levels of crime as well as different trends in crime over the past decade. This richer model is also motivated by fit to the data: a regression model with neighborhood-specific coefficients explains significantly more variation according to an F-test.

That said, we do not expect that every single neighborhood in Philadelphia would have unique coefficients, so we still risk over-parametrization with this model. We address this over-parameterization by imposing shared prior distributions for the neighborhood-specific coefficients from our time trend model (2.5),

$$\boldsymbol{\alpha} \sim \mathrm{N}\left(\alpha_0 \cdot \mathbf{1}\,,\, \tau_\alpha^2 \cdot \mathbf{I}\right) \tag{2.6}$$

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\beta_0 \cdot \mathbf{1}\,,\, \tau_\beta^2 \cdot \mathbf{I}\right) \tag{2.7}$$

$$\boldsymbol{\gamma} \sim \mathrm{N}\left(\mathbf{0}\,,\, \tau_\gamma^2 \cdot \mathbf{I}\right) \tag{2.8}$$

where we denote our collection of neighborhood specific coefficients with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$. $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)$ collects the coefficients (partial effects) of the predictor variables which are shared by all neighborhoods.

We complete this model formulation by placing flat priors on the global means $\alpha_0$ and $\beta_0$,

$p(\alpha_0, \beta_0) \propto 1$, and inverse gamma priors on the variance parameters

$$\sigma^2 \sim \text{Inv-Gamma}(a_\sigma, b_\sigma)$$
$$\tau_\alpha^2 \sim \text{Inv-Gamma}(a_\alpha, b_\alpha)$$
$$\tau_\beta^2 \sim \text{Inv-Gamma}(a_\beta, b_\beta)$$
$$\tau_\gamma^2 \sim \text{Inv-Gamma}(a_\gamma, b_\gamma).$$

The variance hyper-parameters are tuned in an empirical Bayes fashion so that the prior mean of the variance parameters is equal to the variance estimated from the model with no shrinkage, and the prior variance is small. Using non-informative priors for these variance parameters produced nearly identical results. See Section A.4 in the appendix for details.

This Bayesian hierarchical model shares information between neighborhoods by shrinking the neighborhood specific coefficients $\alpha_i$ and $\beta_i$ towards global parameters $(\alpha_0, \beta_0)$ for the entire city. For this reason, we refer to this approach as the *global shrinkage* model.

However, this global shrinkage model does not account for the spatial proximity between neighborhoods when sharing information. We expect close neighborhoods to behave similarly while we want distant neighborhoods to be informative but not as directly influential as adjacent ones. In other words, we may prefer a model that imposes *local* shrinkage rather than global shrinkage.

A model with local sharing of information would also be better able to address the substantial spatial correlation that we see in our application. Testing with Moran's $I$ shows that the residuals from the global shrinkage model are significantly spatially correlated. In the next subsection, we will explore conditional auto-regressive models for local sharing of information.

### 2.3.3. Time Trends with a Spatial Conditional Auto Regressive Model

A popular way of incorporating spatial information is through a prior distribution that is specified according to a Conditional Auto Regressive (CAR) model, which was introduced in its most general formulation by Besag (1974). The CAR model is a Gaussian Markov random field which induces spatial dependence through an adjacency matrix for the areal units, which in our case are neighborhoods in Philadelphia.

Several variations of this CAR framework are reviewed and compared in Lee (2011). In this work, we will use the proper CAR formulation introduced by Leroux et al. (2000).

Let $\boldsymbol{\theta}$ denote a vector of elements that are potentially spatially correlated, such as our neighborhood-specific intercepts $\boldsymbol{\alpha}$ or slopes $\boldsymbol{\beta}$. Leroux et al. (2000) defines the distribution of each $\theta_i$ given the other $\theta_{-i}$ as a normal distribution centered at a weighted average of a global mean and the $\theta_j$'s from bordering neighborhoods,

$$\theta_i \mid \boldsymbol{\theta}_{-i}, \theta_0, \tau^2 \sim \mathrm{N}\left(\frac{\rho \sum_j w_{ij}\,\theta_j + (1-\rho)\,\theta_0}{\rho \sum_j w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho \sum_j w_{ij} + (1-\rho)}\right), \tag{2.9}$$

where $w_{ij}$ are adjacency weights that are equal to 1 if the neighborhoods $i$ and $j$ share a border and equal to 0 otherwise.

We collect these adjacency weights $w_{ij}$ into an adjacency matrix $\mathbf{W}$ that we assume (for now) to be known since we can easily use the shapefiles from the US Census Bureau to determine which of the 1336 neighborhoods (census block groups) share a border.

For now, we consider these adjacency weights $w_{ij}$ to be fixed. However, in Section 2.3.4 we will extend our model to allow those weights to vary since some borders may represent *barriers* between neighborhoods (e.g. highways or rivers), in which case we would not want to share information across that particular border.

The parameter $\rho \in [0, 1]$ represents the strength of the spatial correlation between the components of $\boldsymbol{\theta}$, where larger values of $\rho$ correspond to a stronger influence of bordering

neighborhoods. In the special case of $\rho = 0$, the CAR prior (2.9) reduces to the *global shrinkage* prior (2.6)-(2.7).

It can be proved (Banerjee et al., 2014, Ch.3) using Brook's lemma (Brook, 1964), that the joint distribution of $\boldsymbol{\theta}$ is uniquely determined by the set of conditional distributions defined in 2.9:

$$\boldsymbol{\theta}|\theta_0, \tau^2 \sim \mathrm{N}\left(\theta_0 \cdot \mathbf{1}\,,\, \tau^2 \cdot [\rho(\mathbf{D}_W - \mathbf{W}) + (1-\rho)\mathbf{I}]^{-1}\right) \tag{2.10}$$

where $\mathbf{1}$ is a vector of 1's and $\mathbf{D}_W - \mathbf{W}$ is the Laplacian matrix based on our neighborhood adjacency matrix $\mathbf{W}$. For values of $\rho$ in $[0, 1)$ the joint distribution is proper, while for $\rho = 1$ the distribution is degenerate (Lee, 2011). By adding the constraint $\sum_i(\theta_i - \theta_0) = 0$ we can get a distribution for a $n$-dimensional vector, concentrated in a $(n-1)$-dimensional subspace; this is known as the intrinsic CAR by Besag et al. (1991).

We will employ this CAR model as prior distributions for the vectors of time trend coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We assume $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are *a priori* independent. In vector form, the CAR model (2.9) corresponds to the following prior distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$,

$$\boldsymbol{\alpha} \sim \mathrm{N}\left(\alpha_0 \cdot \mathbf{1}\,,\, \tau_\alpha^2 \cdot \boldsymbol{\Sigma}\right) \tag{2.11}$$

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\beta_0 \cdot \mathbf{1}\,,\, \tau_\beta^2 \cdot \boldsymbol{\Sigma}\right) \tag{2.12}$$

where $\boldsymbol{\Sigma}^{-1} = \rho(\mathbf{D}_W - \mathbf{W}) + (1-\rho)\mathbf{I}$.

We use the same prior distributions for $\alpha_0$ and $\beta_0$ and our variance parameters as in the *global shrinkage* model in the previous subsection. For the additional spatial parameter $\rho$, we choose a Beta$(10, 10)$ prior distribution which has mean equal to 0.5 and a small variance in order to avoid the endpoints of the interval $[0, 1]$.

The posterior distributions for the spatial CAR model and the global shrinkage model (Section 2.3.2) can be implemented via a Gibbs sampler (Geman and Geman, 1984). Implementation details are given in sections A.2 and A.3 in the appendix.

### 2.3.4. Allowing Neighborhood Border Weights to Vary

For most types of areal data, the weights $\mathbf{W}$ that encode the spatial connection between the areal units are considered to be fixed and known. In our data context, the areal units are neighborhoods and the weights $\mathbf{W}$ encode which neighborhoods share a border and hence induce shrinkage on each other in our spatial CAR models outlined in Section 2.3.3.

However, within any large city, some borders between neighborhoods consist of natural or artificial *barriers* such as rivers, highways or train tracks. These barriers could reduce the similarity in crime trends between neighborhoods, and so we would not want to shrink estimates across those barriers. The implication of these barriers for the spatial CAR models in Section 2.3.3 are that some weights $w_{ij} = 1$ should really be $w_{ij} = 0$ since those neighborhoods share a border that is actually a barrier.

Attempting to set which borders should actually be barriers manually would be tedious for a large city and also require extensive domain knowledge and subjective decision making. We instead prefer to infer these barriers from the data by allowing a subset of weights $w_{ij}$ to be random variables in our model.

Specifically, we consider the set of indices of pairs of neighborhoods which share a border according to the geography of Philadelphia. The matrix $\mathbf{W}$ is symmetric so the random variables $w_{ij}$ and $w_{ji}$ are considered to be the same object. We model the $w_{ij}$ for neighborhood pairs that share a border as Bernoulli random variables with an prior probability $\phi$ of $w_{ij} = 1$. Any weights $w_{ij} = 0$ according to the geography of Philadelphia will remain fixed at $w_{ij} = 0$ since we do not want to form connections between non-proximal neighborhoods.

We expect *a priori* that the probability $\phi$ will be close to 1, since relatively few borders between neighborhoods actually should be barriers. For this reason we choose the prior for $\phi$ to be a Beta$(9, 1)$ distribution which has mean close to one and small variance.

Moreover, we expect that the spatial distribution of the neighborhood-specific crime levels

($\alpha_i$) may be different from the neighborhood-specific trends in crime over time ($\beta_i$), so we allow for different barriers when we model the distribution of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In particular, we consider two random matrices $\mathbf{W}^\alpha$ and $\mathbf{W}^\beta$ where a subset of the elements of these matrices are random as described above: $w_{ij}^\alpha \,|\, \phi^\alpha \sim$ Bernoulli($\phi^\alpha$) and $w_{ij}^\beta \,|\, \phi^\beta \sim$ Bernoulli($\phi^\beta$) for neighborhood pairs $(i, j)$ that share a border.

These two weight matrices then determine the local shrinkage of our spatial CAR model from the previous subsection:

$$\boldsymbol{\alpha} \,|\, \mathbf{W}^\alpha \sim \mathrm{N}\left(\alpha_0 \cdot \mathbf{1}\,,\, \tau_\alpha^2 \cdot \boldsymbol{\Sigma}_\alpha\right) \tag{2.13}$$

$$\boldsymbol{\beta} \,|\, \mathbf{W}^\beta \sim \mathrm{N}\left(\beta_0 \cdot \mathbf{1}\,,\, \tau_\beta^2 \cdot \boldsymbol{\Sigma}_\beta\right) \tag{2.14}$$

where $\boldsymbol{\Sigma}_\alpha^{-1} = \rho \cdot (\mathbf{D}_{W^\alpha} - \mathbf{W}^\alpha) + (1-\rho)\mathbf{I}$ and $\boldsymbol{\Sigma}_\beta^{-1} = \rho \cdot (\mathbf{D}_{W^\beta} - \mathbf{W}^\beta) + (1-\rho)\mathbf{I}$.

Allowing variable border weights can lead to over-parametrization since we are adding as many parameters as the number of borders, which makes the shrinkage imposed by prior parameters $\phi^\alpha$ and $\phi^\beta$ important. A more sophisticated approach, which is the focus of ongoing work, would be to partition our areal units into clusters with barriers represented as cluster boundaries.

To implement this extended model with some variable border weights, a step is added to our Gibbs sampler that samples each border weight conditional on the current values of the other model parameters. Details are given in sections A.2 and A.3 in the appendix.

The idea of detecting discontinuities at boundaries is often referred to as *wombling* after the seminal work of Womble (1951) and has been very popular in the disease mapping literature. However, most papers have approached detection of boundaries as a selection problem that is performed after inference (see, e.g. Boots, 2001; Li et al., 2011; Banerjee et al., 2012; Lu and Carlin, 2005; Lee and Mitchell, 2013).

In contrast, we incorporate the possibility of discontinuities at boundaries directly into our

model through variable $\mathbf{W}^{\alpha}$ and $\mathbf{W}^{\beta}$, which allows us to incorporate potential barriers into our estimation of neighborhood-specific parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Lee and Mitchell (2012) and Lu et al. (2007) take a similar approach in the context of disease mapping, but with a more elaborate model for $\mathrm{P}(w_{ij} = 1)$ that is a function of dissimilarity between covariate values in units $i$ and $j$.

In Section 2.4, the different models presented in this section are compared in terms of their accuracy of their in-sample and out-of-sample predictive accuracy. We then visualize the estimated trends in crime over time in Philadelphia and discuss several insights from our results in Section 2.5.

## 2.4. Comparison of Predictive Accuracy

In the previous section, we outlined a no shrinkage model (Section 2.3.2) and several hierarchical Bayesian models for estimating the neighborhood-level trend in crime over time, including a global shrinkage model (Section 2.3.2), a spatial CAR models for local shrinkage (Section 2.3.3), and finally an extension of the spatial CAR model to allow a subset of border weights to vary (Section 2.3.4).

We now compare each of these model alternatives based on several measures of the accuracy of their predictions on both in-sample and out-of-sample hold-out data. Recall that we have 10 years of crime data for the city of Philadelphia, from the beginning of 2006 to the end of 2015. We estimate each model using the crime data for the first nine years (2006-2014).

We assess the *in-sample* accuracy of each model by computing the mean squared error of the predictions of violent crime totals for 2014, which is a year that was included in model estimation,

$$\mathrm{MSE}_{\mathrm{in}} = \frac{1}{1336} \sum_{i=1}^{1336} (y_{i,2014} - \hat{y}_{i,2014})^2. \tag{2.15}$$

We assess the *out-of-sample* accuracy of each model by computing the mean squared error of the predictions of violent crime totals for 2015, which is a year that was *not* included in model estimation,

$$\text{MSE}_{\text{out}} = \frac{1}{1336}\sum_{i=1}^{1336}(y_{i,2015} - \hat{y}_{i,2015})^2. \tag{2.16}$$

To ensure our evaluation is not overly dependent on any idiosyncratic aspects of the 2015 data, we also calculate the *cross-validated out-of-sample* accuracy of each model by calculating the mean square error $\text{MSE}^t_{\text{out}}$ when using year $t$ as the hold out data in the same way that 2015 is used as the hold out data in (2.16), i.e.

$$\text{MSE}_{\text{cv}} = \frac{1}{10}\sum_{t=1}^{10}\text{MSE}^t_{\text{out}} \quad \text{where} \quad \text{MSE}^t_{\text{out}} = \frac{1}{1336}\sum_{i=1}^{1336}(y_{i,t} - \hat{y}_{i,t})^2. \tag{2.17}$$

In Table 1, we compare the predictive accuracy of four different models with neighborhood-specific coefficients outlined in Section 2.3: 1. the time trend model (2.5) without shrinkage between neighborhoods, 2. the global shrinkage model with priors (2.6) and (2.7), 3. the local shrinkage model with spatial CAR priors (2.11) and (2.12) and 4. the local shrinkage spatial CAR model with variable borders (2.13) and (2.14). For additional reference, we also provide the mean square error for fitting a single trend ("Global $\alpha$, $\beta$") across the entire city.

We see in Table 1 that the model with a global trend over time ("Global $\alpha$, $\beta$") for the entire city has very poor predictive accuracy compared to the models that allow neighborhood-specific time trends ("Separate $\alpha_i$, $\beta_i$").

Among the neighborhood-specific time trend models, the global shrinkage model has substantially lower out-of-sample mean square errors than the baseline time trend model without any shrinkage between neighborhoods. The best in-sample mean squared error was achieved by the model without shrinkage, as we expect from the least square method,

| Model | $MSE_{in}$ | $MSE_{out}$ | % change | $MSE_{cv}$ | Moran's $I$ |
|---|---|---|---|---|---|
| Global $\alpha$ , $\beta$ | 0.3558 | 0.3694 | +182.4 | 0.3043 | - |
| Separate $\alpha_i$ , $\beta_i$ Models | | | | | |
| No Shrinkage (2.5) | 0.0567 | 0.1308 | - | 0.1001 | 0.17 |
| Global Shrinkage (2.6)-(2.7) | 0.0698 | 0.1080 | -17.4 | 0.0928 | 0.17 |
| Spatial CAR (2.11)-(2.12) | 0.0703 | 0.1052 | - 19.5 | 0.0922 | 0.61 |
| Variable Borders (2.13)-(2.14) | 0.0706 | 0.1069 | -18.2 | 0.0927 | 0.49 |

Table 1: Comparison of predictive accuracy between the different models outlined in Section 2.3. The mean squared error for both in-sample and out-of-sample predictions are provided, as well as the percentage change in $MSE_{out}$relative to model (2.5) without shrinkage. We also provide the Moran's $I$ measure of spatial correlation calculated on the estimated time trends $\beta_i$ from each model.

though at a cost of having the worst out-of-sample accuracy.

The model with local shrinkage via the spatial CAR prior further reduces the out-of-sample mean square errors compared to the global shrinkage model. The model that allows variable borders does not further improve the out-of-sample mean squared errors, though we explore in Section 2.5.3 that it helps with the interpretation.

Table 1 also provides Moran's $I$ measure of spatial autocorrelation, calculated on the posterior mean of the neighborhood-specific time trends ($\beta_i$'s). We see that the spatial CAR model induces a larger spatial correlation in the $\beta_i$'s than the models with global shrinkage or without shrinkage. The local shrinkage model has a Moran's I value of 0.61 ($s.e.$ = 0.016), which suggests there is substantial spatial autocorrelation in the change in crime within Philadelphia.

In summary, allowing for local shrinkage of the neighborhood-specific crime trend coefficients via the spatial CAR priors (2.11) and (2.12) leads to the best out-of-sample predictive accuracy. In Section 2.5, we visualize the parameters of this model and discuss the implications of these results for crime in Philadelphia.

Although the variable border model extension does not improve out-of-sample predictive accuracy, we will also see in Section 2.5 that visualizing the borders that have been turned into barriers by this model provide insight into discontinuities in crime trends in the city of

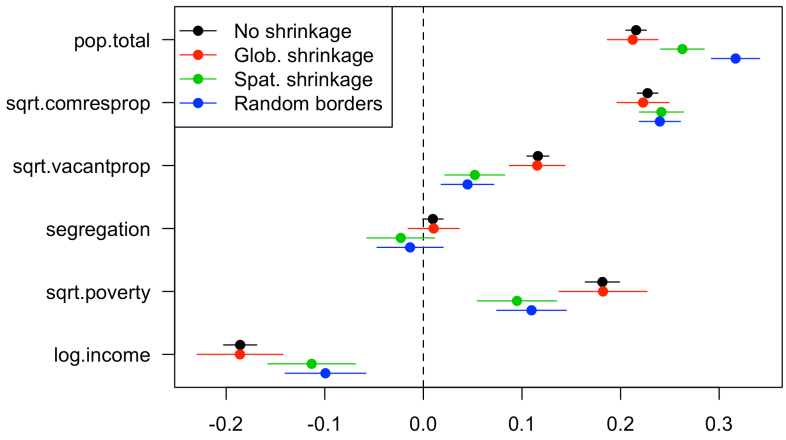


Figure 2: Estimated partial effects $\gamma_d$ from four different models indicated in the legend. For the no shrinkage model, we plot the maximum likelihood estimate and 95% confidence interval. For three Bayesian shrinkage models, we plot the posterior mean and 95% posterior interval.

Philadelphia.

## 2.5. Interpretation of Model Parameters

In Section 2.5.1, we examine the estimated partial effects for the static predictor variables created from the data outlined in Section 2.2. We then visualize and compare the estimated neighborhood-specific levels ($\alpha_i$'s) and time trends ($\beta_i$'s) on crime from our different models in Section 2.5.2. In Section 2.5.3, we examine the results from our model extension outlined in Section 2.3.4 that allows a subset of neighborhood borders in Philadelphia to be estimated as barriers. Finally, in Section 2.5.4 we discuss the neighborhoods with the most extreme levels and changes in crime over time over the past ten years in Philadelphia.

### 2.5.1. Partial Effects of Static Predictors

Figure 2 gives the estimated partial effects $\gamma_d$ for each static predictor variable $d$ from the four models outlined in Section 2.3. We provide additional numerical details in Table S1 of appendix A.

We see that among the six predictor variables created in Section 2.2, only the segregation

measure is not a significant predictor of crime. All predictor variables are on the same scale and so we can directly compare the values of their partial effects.

We see that the strongest predictors of crime are total population and the commercial versus residential proportion, with more populated and more commercial neighborhoods being associated with higher crime. Income and poverty are also significantly predictive of violent crimes but we must be more cautious about interpreting these partial effects given the high collinearity between income and poverty. Each of these observations on the partial effects $\boldsymbol{\gamma}$ is relatively consistent across the four models outlined in Section 2.3.

### 2.5.2. Visualizing Neighborhood-Specific Coefficients

Our primary interest in terms of interpretation are the estimated neighborhood-specific coefficients, $\alpha_i$'s and $\beta_i$'s, that represent the level of violent crimes and change in violent crimes over time in Philadelphia, respectively.

In Figures 3 and 4, we give maps where each block group in Philadelphia is colored by the estimated neighborhood-specific levels of crime $\hat{\alpha}_i$ and changes in crime over time $\hat{\beta}_i$ respectively, from the four models outlined in Section 2.3. We see substantial heterogeneity between neighborhoods in Philadelphia, both in terms of the their estimated crime levels ($\hat{\alpha}_i$'s) and changes in crime over time ($\hat{\beta}_i$'s). Regardless of the model, most neighborhoods in the city show decreasing trends in crime over time (negative $\beta$'s) with a small subset of neighborhoods showing an increasing trend.

The shrinkage imposed by the global shrinkage model is more visually striking for the change in violent crime over time than the overall level of crime. The maps of the $\hat{\alpha}_i$'s from the no shrinkage and global shrinkage models are almost indistinguishable in Figure 3 whereas the map of the $\hat{\beta}_i$'s from the global shrinkage model has been shifted substantially compared to the no shrinkage map in Figure 4. This observation suggests that there is more substantial heterogeneity between neighborhoods in terms of their overall level of crime compared to their change in crime over time.

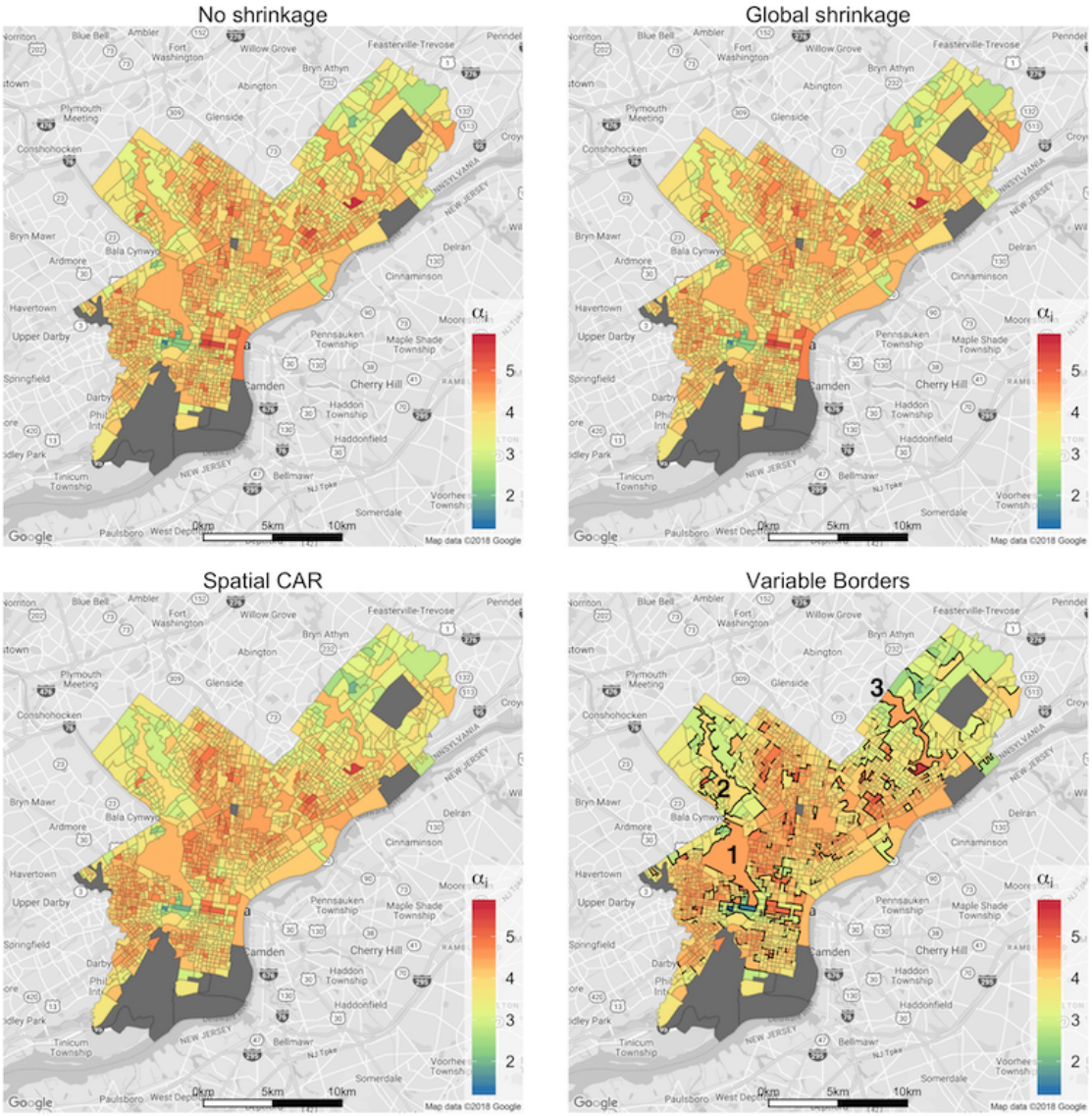


Figure 3: Maps of Philadelphia colored by the estimated intercept from our four different models. **Top-left:** Maximum likelihood estimates of $\alpha_i$ from the no shrinkage model (2.5). **Top-right:** Posterior means of $\alpha_i$ from the global shrinkage model (2.6)-(2.7). **Bottom-left:** Posterior means of $\alpha_i$ from the spatial CAR model (2.11)-(2.12). **Bottom-right:** Posterior means of $\alpha_i$ from the spatial CAR model with variable borders (2.13)- (2.14). The black lines represent borders turned into barriers. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

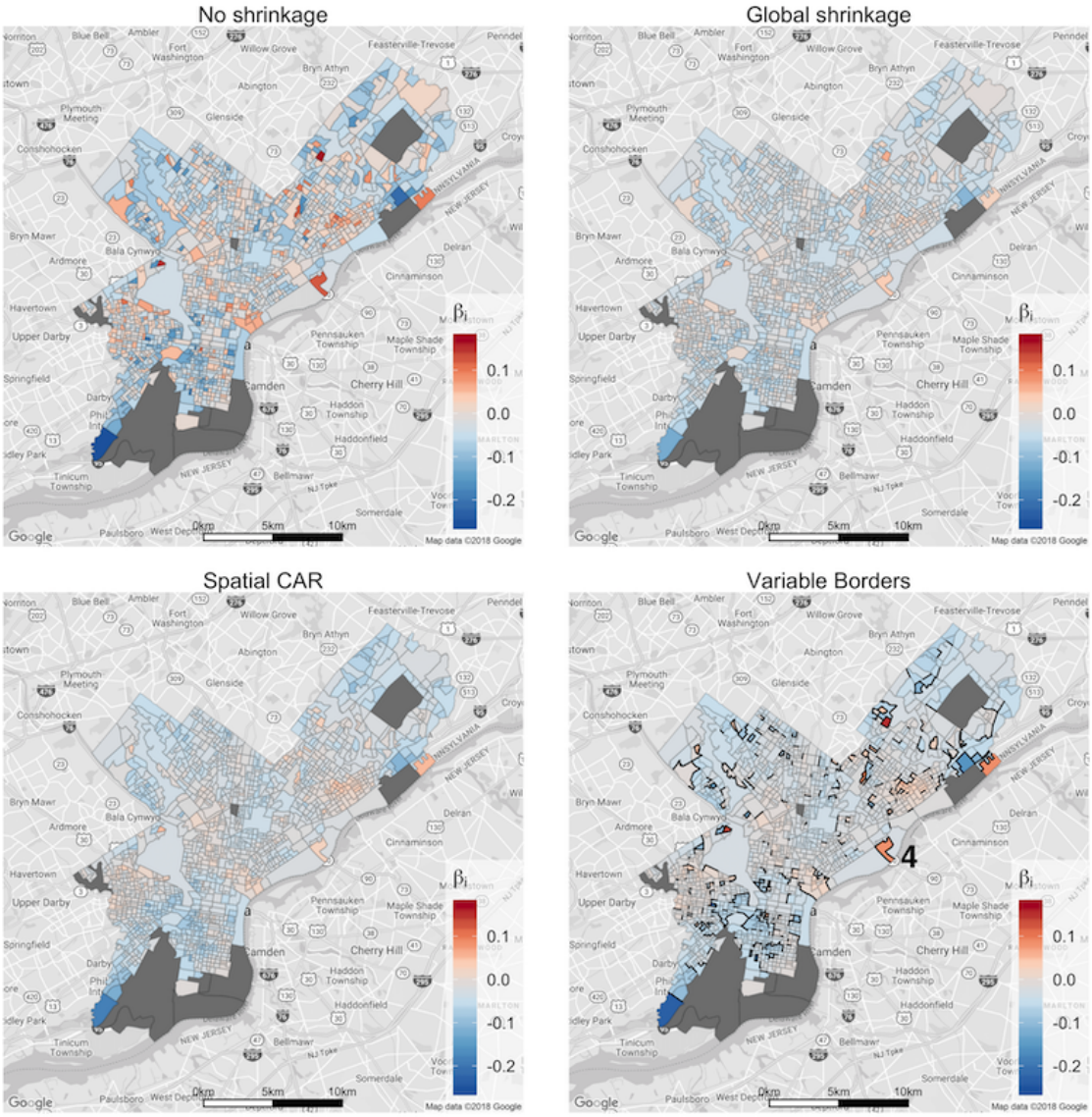


Figure 4: Maps of Philadelphia colored by the estimated slope on time from our four different models. **Top-left:** Maximum likelihood estimates of $\beta_i$ from the no shrinkage model (2.5). **Top-right:** Posterior means of $\beta_i$ from the global shrinkage model (2.6)-(2.7). **Bottom-left:** Posterior means of $\beta_i$ from the spatial CAR model (2.11)-(2.12). **Bottom-right:** Posterior means of $\beta_i$ from the spatial CAR model with variable borders (2.13)-(2.14). The black lines represent borders turned into barriers. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

This heterogeneity in the mean level of crime is expected as it is influenced by many years of transformation in the city of Philadelphia that led to its current built and social environment. Differences in these overall spatial crime patterns can be addressed by urban planners, whose effects are long-lasting (Johnson et al., 2008). In contrast, differences in the trend over time identify shorter-term patterns, which can be addressed with interventions by local police departments.

The overall level of crime also seems to have a greater inherent spatial correlation between proximal neighborhoods than the change in crime over time. The Moran's I values calculated from the estimated $\hat{\alpha}_i$'s are I = 0.33 for both the no shrinkage and global shrinkage models, compared to the value of I = 0.17 from the estimated $\hat{\beta}_i$'s for those same models in Table 1. This is clear also from the maps from the no shrinkage model (top left) in Figures 3 and 4: the estimated $\hat{\beta}_i$'s are more "spotty" and less smooth than the corresponding map of the $\hat{\alpha}_i$'s.

However, once we build spatial correlation into our model via the spatial CAR prior (2.11)-(2.12), the resulting $\hat{\beta}_i$'s are more spatially correlated than the resulting $\hat{\alpha}_i$'s, as can be seen in the lower left of Figures 3 and Figures 4 as well as the corresponding Moran's I = 0.53 for the $\hat{\alpha}_i$'s versus I = 0.61 for the $\hat{\beta}_i$'s. Note that all these reported Moran's I values have a standard error approximately equal to 0.016, and so they are all significantly different from the null hypothesis of no spatial autocorrelation.

Although the smoother maps from the spatial CAR model (lower left of Figures 3 and 4) ease interpretation by identifying larger regions of the city with similar crime dynamics, there is the potential to over-shrink certain neighborhoods that should actually stand out from their neighbors. In any large city, natural or artificial *barriers* such as rivers, highways or rail lines create discontinuities between neighborhoods which should not be smoothed over. In Section 2.5.3, we examine the results from our model extension that allows a subset of borders between neighborhoods to be turned into barriers.

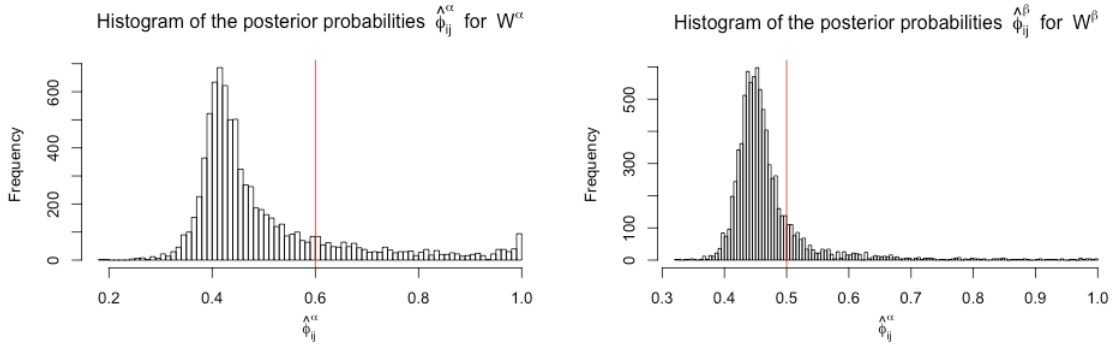


Figure 5: Histograms of the posterior probabilities of each border being turned into a barrier. **Left:** Probabilities for barriers for the $\alpha_i$'s; the threshold to identify the borders turned into barriers is 0.6 (red line). **Left:** Probabilities for barriers for the $\beta_i$'s; the threshold to identify the borders turned into barriers is 0.5 (red line).

### 2.5.3. Borders turned into Barriers

In Section 2.3.4, we extended the spatial CAR model to allow a subset of the weights $w_{ij}$ to vary, which allows the *borders* ($w_{ij} = 1$) between some neighborhoods to be changed into *barriers* ($w_{ij} = 0$); the latter prevent shrinkage between two bordering neighborhoods. Our model has separate weight matrices $\mathbf{W}^{\alpha}$ and $\mathbf{W}^{\beta}$, so a particular border can be turned into a barrier either for the level of crime ($\alpha_i$'s) or the change in crime over time ($\beta_i$'s) or both. Using this model, we estimate the posterior probability that we change a border into a barrier for each border between proximal neighborhoods in Philadelphia.

Figure 5 gives the distribution of the estimated posterior probability of a border being turned into barrier for each border encoded in the weight matrices $\mathbf{W}^{\alpha}$ and $\mathbf{W}^{\beta}$. These distributions seem to have two components: a main mode representing the behavior of the majority of the borders, which has a low probability of being turned into a barrier, and a "tail" component which has a higher probability of being turned into a border.

It is clear that many more borders have a high probability of being a barrier for the level of crime ($\alpha_i$'s) compared to the change in crime over time ($\beta_i$'s). In other words, our variable border model is detecting more discontinuities between bordering neighborhoods in the level

of crime compared to the change in crime over time. In Section A.5.2 of the appendix, we explore an alternative model that only allows variable borders for the mean level of crime.

In the lower right panels of Figures 3 and 4, we provide maps of Philadelphia where we have highlighted any borders between neighborhoods that have been inferred by our model to have a high probability of being barriers. These particular highlights are based on posterior probabilities larger than 60% for $\mathbf{W}^{\alpha}$ and larger than 50% for $\mathbf{W}^{\beta}$.

We see in the lower right panel of Figure 3 that barriers have been detected around several parks including Fairmount Park, Wissahickon Valley Park, and Pennypack Creek Park (indicated by the black numbers 1, 2 and 3 respectively in the lower right panel of Figure 3). In these cases, our model has automatically detected several natural geographic structures within Philadelphia as locations which have discontinuities in the level of crime.

We also see that some estimated barriers have isolated particular neighborhoods from their proximal neighbors. For example, the neighborhood of Bridesburg (indicated by the black number 4 in the lower right panel of Figure 4) seems to have a much more positive trend on crime over time than its surrounding neighborhoods.

As barriers highlight the boundaries of regions that display differences in either in the level of crime or the trend in crime over time, these barriers can be used by police departments and city planners for delineating the possible limits of effectiveness for interventions or as potential targets for interventions themselves.

### 2.5.4. Neighborhoods with Most Extreme Crime Trends

To further understand which regions of Philadelphia have the most extreme levels of crime and trends in crime over time, we can examine the most extreme intercepts ($\alpha_i$'s) and slopes ($\beta_i$'s) found by our fitted models. Specifically, we focus on the estimated $\alpha_i$'s and $\beta_i$'s from the local shrinkage spatial CAR model (2.11)-(2.12) that had the best out-of-sample predictive performance in Table 1.

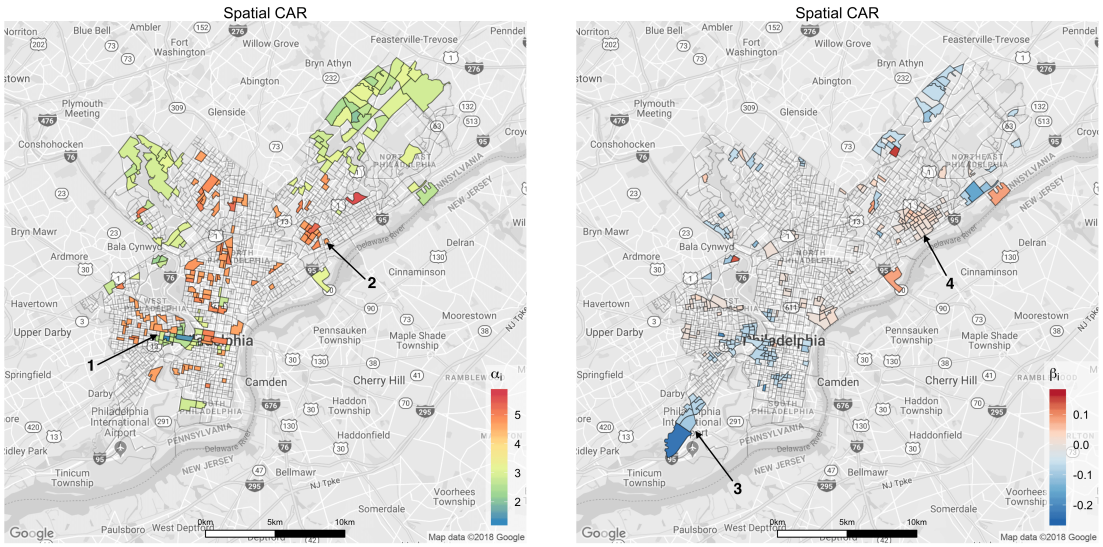


Figure 6: **Top:** The 50 neighborhoods with the largest $\hat{\alpha}_i$'s (red) and 50 neighborhoods with the smallest $\hat{\alpha}_i$'s (green). **Bottom:** The 50 neighborhoods with the largest $\hat{\beta}_i$'s and 50 neighborhoods with the smallest $\hat{\beta}_i$'s. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

Figure 6 provides maps that highlight the most extreme (largest 50 and smallest 50) neighborhoods in terms of the estimated level of crime ($\hat{\alpha}_i$'s) and in terms of the estimated change in crime over time ($\hat{\beta}_i$'s).

We see that the region of University City in West Philadelphia (black number 1 in the top panel of Figure 6) is an interesting transitional area that contains both neighborhoods with the highest and lowest levels of crime in the city. We also see that the area of Frankford (black number 2 in the top panel of Figure 6) has neighborhoods with high levels of crime. This area is a major transportation hub for the Northeast region of Philadelphia.

The SW region of Philadelphia, specifically the Elmwood and Eastwick neighborhoods (black number 3 in the bottom panel of Figure 6) have seen some of the largest reductions in crime over the past decade in Philadelphia. We also see some regions of the city that are showing increases in crime over that same time period, such as the Wissinoming and Tacony neighborhoods (black number 4 in the bottom panel of Figure 6) that are just to the northeast of the high crime neighborhoods of Frankford (black number 2 in the top

panel of Figure 6).

In Section A.5 of the appendix, we provide additional visualizations of the neighborhood-specific parameters that are significantly different from the overall mean across the city as well as the widths of the credible intervals for these parameters.

## 2.6. Discussion

Reliable estimation of the change in crime over time at the local neighborhood level is a crucial step towards a better understanding of the determinants of public safety in large urban areas. With a focus on the city of Philadelphia, we have explored several Bayesian approaches to modeling crime trends within the areal units of neighborhoods while sharing information either globally or locally across the city.

Imposing local shrinkage between proximal neighborhoods via a spatial conditional autoregressive (CAR) prior gives the best out-of-sample predictions of violent crime compared to models that impose global shrinkage or no shrinkage at all between neighborhoods. We also explore allowing the weight matrix of our spatial CAR model to vary in order to detect neighborhood borders that represent spatial discontinuities in the level of crime or change in crime over time. In this way, we automatically detect several natural barriers in the geography of Philadelphia. Our model estimates also identify the regions of Philadelphia with the most extreme levels of violent crime as well as the largest increases and reductions in crime over the period of 2006-2015.

# Chapter 3

# Bayesian Clustering with Particle Optimization

## 3.1. Introduction

Accurate modeling of urban crime dynamics benefits many constituents: law enforcement officials can make more informed decisions about how to deploy resources to ensure public safety, urban planners can better understand how socio-economic factors and the built environment affect crime, and city officials can develop community programs and interventions to improve the overall quality of life in the city. In this work, we study how crime has evolved in the city of Philadelphia between 2006 and 2017 with a focus on finding clusters of neighborhoods with similar crime dynamics.

Bayesian hierarchical modeling is a very natural way to study crime at the neighborhood level as it allows us to "borrow strength" between spatially adjacent neighborhoods. In fact, Balocchi and Jensen (2019) have demonstrated that Bayesian models that encourage spatial shrinkage can yield more accurate predictions than models that do not introduce dependencies between parameters from adjacent neighborhoods. Following that work, we propose a model that extends Bernardinelli et al. (1995)'s linear model to crime incidents with spatially varying intercepts (mean level of crime) and spatially varying slopes (time trend).

Priors based on conditionally auto-regressive (CAR) models (Besag, 1974) are workhorses in the Bayesian spatial statistics literature that encourage shrinking each neighborhood's parameters towards the average value of the parameters from adjacent neighborhoods. Though these models are an intuitive and popular way to "share information" between spatially

Adapted from a research article:
**Balocchi, C.**, Deshpande, S. K., George E. I. and Jensen, S. T. (2019) "Bayesian Spatial Clustering of Crime in Philadelphia with Particle Optimization" *arXiv 1912.00111*

adjacent regions, they can introduce a level of smoothness at odds with the realities of complex urban environments. In fact, as we will see in Section 3.2, while crime incidents in Philadelphia display considerable spatial correlation, there are also many sharp spatial discontinuities. This is because geographic aspects of the city, such as major streets, parks, and rivers, and latent socioeconomic divisions can create barriers that may be associated with discontinuities in crime patterns.

In the context of crime modeling, using a CAR prior without accounting for potential discontinuities can lead to poor estimation of crime around these geographic or socioeconomic barriers. Although manually adjusting the CAR prior to prevent smoothing over these boundaries is conceptually simple, it presupposes knowledge about the location of these discontinuities, which are often latent or unknown. A far more elegant and agnostic approach is to use the data itself to identify the discontinuities.

There is a very rich literature on data-adaptive strategies for detecting discontinuities at the border between adjacent neighborhoods, also known as *wombling*. One approach to wombling involves first fitting a simple model that does not account for potential discontinuities and then identifying jumps in the fitted values (see, e.g., Boots (2001), Li et al. (2011), Banerjee et al. (2012), Lu and Carlin (2005), and Lee and Mitchell (2013)). Alternatively, many authors directly model uncertainty about which borders correspond to sharp discontinuities within larger Bayesian hierarchical models (see, e.g., Lee and Mitchell (2012), Lu et al. (2007), and Balocchi and Jensen (2019)). While directly modeling the uncertainty in discontinuity locations is intuitively appealing, these latter models are heavily over-parametrized; in fact, they introduce one latent parameter for each pair of adjacent neighborhoods.

Rather than look for individual discontinuities between pairs of neighborhoods, we instead aim to identify *clusters* of neighborhoods that exhibit similar crime dynamics. Compared to wombling, clustering encourages dimensionality reduction while maintaining model interpretability and flexibility. In this work, we propose a "CAR–within–clusters" model

where we introduce *two latent spatial partitions* of neighborhoods in Philadelphia, one for the mean levels of crime and one for the temporal trends. We then specify separate CAR priors on the neighborhood-specific parameters within each cluster of each partition. We describe our data and introduce this model in Section 3.2.

Like similar spatial clustering approaches (see, e.g., Knorr-Held and Raßer (2000), Denison and Holmes (2001), Feng et al. (2016), and references therein), we treat parameters arising from different clusters independently *a priori*. However, unlike these works, we do not assume that all parameters within a cluster are equal. Instead, we allow the parameters to vary smoothly within each cluster. Our approach combines positive aspects of clustering and wombling: we are able to find areas displaying different crime dynamics and simultaneously interpret borders between clusters as barriers corresponding to spatial discontinuities.

In our implementation, we have three primary tasks: (i) identify the two underlying spatial partitions, (ii) estimate the neighborhood-level parameters, and (iii) make predictions of future crime incidents while accounting for our uncertainty about the partitions. These goals are complicated by the combinatorial vastness of the latent product space of spatial partitions, rendering typical stochastic search techniques computationally prohibitive. We instead focus on posterior optimization. However, rather than simply finding the *maximum a posteriori* (MAP) partitions, we propose an extension of Ročková (2018)'s ensemble optimization framework that simultaneously identifies multiple partitions with high posterior probability by solving a *single* optimization problem. In Section 3.3, we show that solving this problem is formally equivalent to finding a particular variational approximation of the discrete posterior distribution of the pairs of partitions. We introduce a new local search strategy that, at a high level, runs several greedy searches that are made "mutually aware" by an entropy penalty. This penalty promotes diversity among the estimated partitions by discouraging different search paths from visiting the same point in the latent discrete space. By identifying several high posterior probability partitions we can easily incorporate uncertainty about the latent clusterings into our estimation of the pa-

rameters and prediction, with Bayesian Model Averaging (BMA; Raftery et al. (1997)). In Section 3.4, we illustrate our proposed methodology on simulated data before applying it to the Philadelphia data in Section 3.5. We conclude with a discussion of our results and an outline of potential future directions in Section 3.6. A software implementation of our method and all code and data to replicate the results in this work are available at `github.com/cecilia-balocchi/particle-optimization`.

## 3.2. Data and the "CAR–within–clusters" Model

For the first time in decades, Philadelphia is experiencing population growth and its built environment is rapidly evolving; this transformation makes it an interesting real-time case study for examining how crime evolves over time. Our crime data comes from `opendataphilly.org`, where the Philadelphia Police Department publicly releases the location, time, and type of each reported crime in the city. While there has been an overall decrease in the total amount of crime in the city over the last decade, we can obtain a more nuanced understanding by examining the temporal trends at a local neighborhood level. Our analysis focuses on *violent* crimes, which include homicides, rapes, robberies, and aggravated assaults (FBI), aggregated at the census tract level. In all, Philadelphia is divided into $N = 384$ census tracts, which we treat as large neighborhoods in our analysis, as census tracts contain approximately 4,000 inhabitants each.

For the years between 2006 ($t = 0$) and 2017 ($t = 11$), let $c_{i,t}$ be the total number of violent crimes reported in tract $i$ during year $t$. The distribution of crime counts $c_{i,t}$ displays considerably skewness. Similar to Balocchi and Jensen (2019), rather than modeling $c_{i,t}$ directly, we work with an inverse hyperbolic sine transformation (Burbidge et al., 1988a) of the violent crime counts:

$$y_{i,t} = \log\left(c_{i,t} + (c_{i,t}^2 + 1)^{1/2}\right) - \log(2).$$

This transformation is a close approximation of $\log(c_{i,t})$ but is also well-defined for neigh-

borhoods that had a crime count of zero in certain years.

### 3.2.1. Model

To study the crime dynamics at the neighborhood level in Philadelphia, we consider a simple linear regression model:

$$y_{i,t} = \alpha_i + \beta_i(t - \bar{t}) + \varepsilon_{i,t}; \quad \varepsilon_{i,t} \sim N(0, \sigma^2) \tag{3.1}$$

where time $t$ has been centered, so that the parameters $\alpha_i$ and $\beta_i$ respectively represent the mean level of crime and the trend over time of crime in census tract $i$. Linear models are typically employed when the number of time points is small or moderate (Bernardinelli et al., 1995; Anderson et al., 2017).

We can obtain an initial estimate of the average levels $\alpha_i$ and time trends $\beta_i$ of crime by treating each neighborhood independently and computing the maximum likelihood estimates (MLEs) within each neighborhood. Figure 7 displays these estimates and reveals that the broad negative time trend in crime is not uniform across the city. In fact, in a small number of neighborhoods, crime has actually increased over the last decade.

We also see in Figure 7 that, with few notable exceptions, spatially adjacent neighborhoods tends to have similar MLEs, suggesting a high degree of spatial correlation in the neighborhood-level crime dynamics. We take a hierarchical Bayesian approach in order to "borrow strength" between neighborhoods that involves specifying a prior distribution on the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$. Because we expect the tract-specific parameters to display some spatial continuity, we use priors that explicitly introduce dependence between parameters from neighboring tracts.

Conditionally autoregressive (CAR) models are a popular class of such priors and we use a version introduced in Leroux et al. (2000). Letting $W = (w_{i,j})$ be a binary adjacency matrix with $w_{i,j} = 1$ if and only if neighborhoods $i$ and $j$ share a border, we say that the

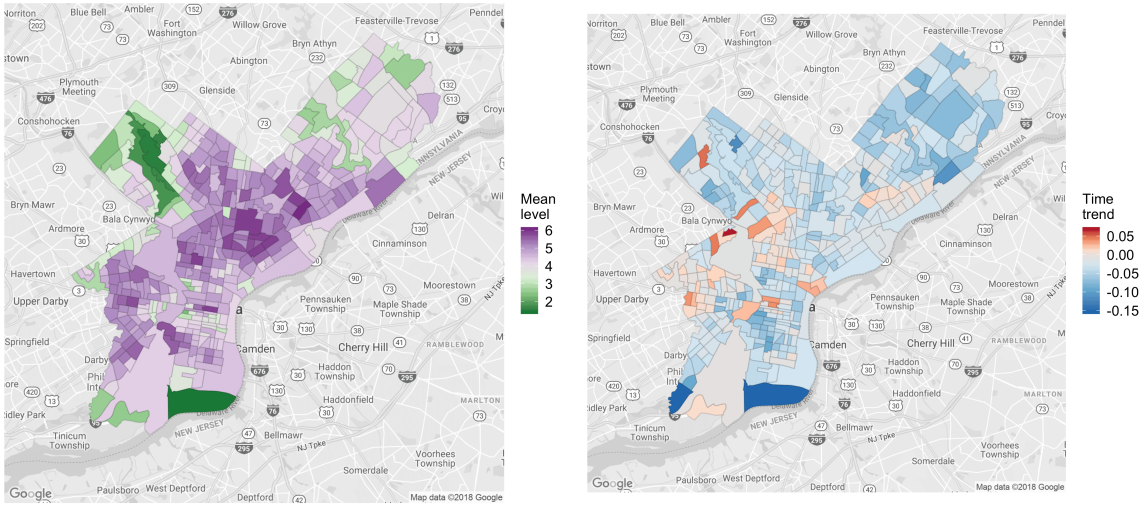


Figure 7: Visualization of the maximum likelihood estimates of the tract-level intercepts $\boldsymbol{\alpha}$ (left panel) and time-trends $\boldsymbol{\beta}$ (right panel) for the model defined in Section 3.2.1

vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ follows a CAR model with grand mean $\overline{\theta}$ and variance scale $\tau^2$ if and only if all of the full conditional distributions have the form

$$\theta_i \mid \boldsymbol{\theta}_{-i}, \overline{\theta}, \tau^2 \sim N\left(\frac{(1-\rho)\overline{\theta} + \rho \sum_j w_{i,j}\theta_j}{1 - \rho + \rho \sum_j w_{i,j}}, \frac{\tau^2}{1 - \rho + \rho \sum_j w_{i,j}}\right).$$

In this CAR model, the conditional mean of $\theta_i \mid \boldsymbol{\theta}_{-i}$ is a weighted average of the grand mean $\overline{\theta}$ and the average of the $\theta_j$'s from the neighborhoods that border neighborhood $i$. The degree to which $\theta_i$ is shrunk toward either of these targets is governed by a parameter $\rho$, which is typically set by the analyst, and the number of neighbors. These full conditionals uniquely determine the joint distribution $\boldsymbol{\theta} \sim N(\overline{\theta}\mathbf{1}_n, \tau^2\Sigma_{\text{CAR}})$ where

$$\Sigma_{\text{CAR}} = \begin{cases} [\rho W^\star + (1-\rho) I_n]^{-1} & \text{if } n \geq 2 \\ \frac{1}{1-\rho} & \text{if } n = 1 \end{cases},$$

$\mathbf{1}_n$ is the $n$-vector of ones, and $W^\star$ is the unweighted graph Laplacian of the adjacency matrix $W$. For compactness, we will write $\boldsymbol{\theta} \mid \overline{\theta}, \tau^2 \sim \text{CAR}(\overline{\theta}, \tau^2, W)$.

However, cities typically contain many geographic and social barriers like rivers and highways that manifest in sharp spatial discontinuities. In the presence of these discontinuities, a naively specified CAR model can induce a level of spatial smoothness among the parameters at odds with the data. To avoid this behavior, we seek *clusters* of parameters that demonstrate considerable spatial continuity within but not between clusters. We introduce two latent partitions of $[N]$, $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$, where $\gamma^{(\cdot)} = \{S_1^{(\cdot)}, \ldots, S_{K^{(\cdot)}}^{(\cdot)}\}$. We refer to the sets $S_k^{(\cdot)}$ as *clusters* and restrict attention to partitions consisting of clusters of spatially connected neighborhoods. We denote the set of all such partitions by $\mathcal{SP}$ and let $\boldsymbol{\gamma} := (\gamma^{(\alpha)}, \gamma^{(\beta)})$ be the pair of latent spatial partitions underlying the mean level of crime and the time trend of crime across all neighorhoods. In what follows, we will simply refer to $\boldsymbol{\gamma}$ as a *particle*.

To simplify our presentation, we describe only the prior over the mean levels of crime $\boldsymbol{\alpha}$; we place an analogous prior on the time trends $\boldsymbol{\beta}$. We place independent CAR priors on the collections $\boldsymbol{\alpha}_k = \{\alpha_i : i \in S_k^{(\alpha)}\}$, so that the joint prior density $\pi(\boldsymbol{\alpha} \mid \gamma^{(\alpha)}, \sigma^2)$ factorizes over the collection of all clusters: $\pi(\boldsymbol{\alpha} \mid \gamma^{(\alpha)}, \sigma^2) = \prod_{k=1}^{K^{(\alpha)}} \pi(\boldsymbol{\alpha}_k \mid \sigma^2)$. To this end, we introduce a collection of grand cluster means $\overline{\boldsymbol{\alpha}} = \{\overline{\alpha}_1, \ldots, \overline{\alpha}_{K^{(\alpha)}}\}$ and model $\boldsymbol{\alpha}_k \mid \overline{\alpha}_k, \sigma^2 \sim \text{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)})$, where $W_k^{(\alpha)}$ is the sub-matrix of $W$ whose rows and columns are indexed by the cluster $S_k^{(\alpha)}$. We further place independent $N(0, a_2\sigma^2)$ priors on the grand cluster means $\overline{\alpha}_k$ and place a fully-specific prior $\Pi_\gamma$ on $\gamma^{(\alpha)}$. In Sections 3.4 and 3.5, we consider two different priors for the latent partitions. The first is a truncated Ewens-Pitman prior with probability mass function

$$\pi(\gamma) \propto \eta^K \prod_{k=1}^{K} (n_k - 1)! \times \mathbf{1}(\gamma \in \mathcal{SP}). \tag{3.2}$$

The second is a truncated uniform prior that assigns equal prior probability to each $\gamma \in \mathcal{SP}$. We note here, however, that the computational strategy introduced in Section 3.3 will work for general priors. We complete our hierarchical prior with an Inverse Gamma prior on the residual variance $\sigma^2 \sim \text{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right)$.

To summarize, our model is

$$
\begin{aligned}
\gamma^{(\alpha)}, \gamma^{(\beta)} &\overset{iid}{\sim} \Pi_\gamma \\
\sigma^2 &\sim \text{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
\overline{\alpha}_1, \ldots, \overline{\alpha}_{K_\alpha} \mid \gamma^{(\alpha)}, \sigma^2 &\overset{iid}{\sim} N(0, a_2\sigma^2) \\
\overline{\beta}_1, \ldots, \overline{\beta}_{K_\beta} \mid \gamma^{(\beta)}, \sigma^2 &\overset{iid}{\sim} N(0, b_2\sigma^2) \\
\boldsymbol{\alpha}_k \mid \overline{\alpha}_k, \sigma^2, \gamma^{(\alpha)} &\sim \text{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)}) \quad \text{for } k = 1, \ldots, K_\alpha \\
\boldsymbol{\beta}_{k'} \mid \overline{\beta}_{k'}, \sigma^2, \gamma^{(\beta)} &\sim \text{CAR}(\overline{\beta}_{k'}, b_1\sigma^2, W_{k'}^{(\beta)}) \quad \text{for } k' = 1, \ldots, K_\beta \\
y_{i,t} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 &\sim N(\alpha_i + \beta_i(t - \bar{t}), \sigma^2)
\end{aligned}
\tag{3.3}
$$

The high degree of conditional conjugacy in (3.3) enables us to derive analytic expressions for quantities such as the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ as well as the conditional posterior expectations $\mathbb{E}[\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{y}]$. The availability of these expressions will be crucial for the posterior exploration strategy we develop below.

Given the residual variance $\sigma^2$ and latent partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$, parameters in different clusters are conditionally independent. In other words, our model falls with the class of conditional product partition models (PPMs) that have been widely used in Bayesian spatial statistics (see, e.g., Knorr-Held and Raßer (2000), Denison and Holmes (2001), and Feng et al. (2016)). Unlike these papers, however, we are interested in recovering two latent partitions, one each for the mean levels and time-trends within each census tract. In this way, our model is similar to Anderson et al. (2017), who also seek two distinct partitions of the set of neighborhoods. However, unlike Anderson et al. (2017), who limit attention to partitions containing five or fewer clusters for computational simplicity, we do not need to impose any *a priori* restriction on the number of clusters.

## 3.3. Posterior Exploration and Summarization

Recall that we have three simultaneous tasks: (i) identify promising particles $\boldsymbol{\gamma} = (\gamma^{(\alpha)}, \gamma^{(\beta)})$, (ii) estimate the mean-levels $\boldsymbol{\alpha}$ and time trends $\boldsymbol{\beta}$ of crime in each neighborhood, and (iii) make predictions about future incidents of crime in each neighborhood. These latter two tasks can generally be expressed as evaluating posterior expectations $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$ where $f$ is any functional of interest. The combinatorial vastness of the space $\mathcal{SP}^2$, which contains all possible pairs of partitions, renders it impossible to enumerate all particles for even small values of $N$. As a result, we cannot compute the posterior probability $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ exactly.

It is tempting to resort to Markov Chain Monte Carlo (MCMC) simulations to approximate expectations $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. We could, for instance, proceed in a Gibbs fashion, alternating between updating the two partitions in each $\boldsymbol{\gamma}$ and updating continuous parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$, while holding the rest fixed. Unfortunately, because we must explore a vast space of pairs of partitions, such MCMC simulations may require a prohibitive amount of time to mix. To get around this difficulty, Anderson et al. (2017) arbitrarily restricted attention to partitions with no more than three to five clusters each. Even with such a restriction, which we will not impose, it is still quite difficult to distill the thousands of resulting draws of $\boldsymbol{\gamma}$ into a single point estimate and to quantify parameter and partition uncertainty.

A popular alternative approach is posterior optimization, which usually focuses on identifying the *maximum a posteriori* (MAP) particle $\hat{\boldsymbol{\gamma}}_{MAP}$ or some other decision-theoretic optimal point estimate (see, e.g., Lau and Green (2007)). One then estimates the marginal expectation $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$ with a "plug-in" estimator $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}_{MAP}]$. Though this procedure might be substantially faster than MCMC, especially if the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ possesses certain ordering properties (Dahl, 2009), it completely eschews exploration of the uncertainty about $\boldsymbol{\gamma}$. As a result, the natural "plug-in" estimator $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}_{MAP}]$ may result in over-confident inference about the function $f$.

Notice, however, that this plug-in estimator may be viewed as a particular instantiation of Bayesian Model Averaging (BMA) (Raftery et al., 1997; Hoeting et al., 1999). At a very high-level, BMA aims to approximate the full marginal expectation

$$\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}] = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \boldsymbol{\gamma}],$$

by first identifying some small subset $\Gamma$ of models and then evaluating the more manageable sum

$$f_\Gamma = \sum_{\boldsymbol{\gamma} \in \Gamma} \pi_\Gamma(\boldsymbol{\gamma} \mid \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}, \boldsymbol{\gamma}],$$

where $\pi_\Gamma$ is the restriction of the posterior $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ to the set $\Gamma$.

Intuitively, the better the restricted posterior $\pi_\Gamma$ approximates the full posterior $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$, the closer $f_\Gamma$ will be to the targeted marginal expectation $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. So rather than just using the top $\boldsymbol{\gamma}$, a natural extension of the MAP plug-in is to use the top $L > 1$ $\boldsymbol{\gamma}$'s. Specifically if we let $\Gamma_L = \{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\}$ be the $L$ particles with largest posterior mass, we consider

$$f_L = \sum_{\ell=1}^{L} \tilde{\pi}(\boldsymbol{\gamma}^{(\ell)} | \mathbf{y}) \mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \boldsymbol{\gamma}^{(\ell)}, \mathbf{y}],$$

where $\tilde{\pi}(\cdot|\mathbf{y})$ is the truncation of $\pi(\boldsymbol{\gamma}|\mathbf{y})$ to $\Gamma_L$. In contrast to the MAP plug-in estimator, $f_L$ averages over more of the particle selection uncertainty and we might reasonably expect it to be a better approximation of the marginal posterior mean $\mathbb{E}[f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mid \mathbf{y}]$. Of course, in order to compute $f_L$ exactly, we know which $L$ particles have the most posterior probability. In the next subsection, we introduce a general strategy for identifying $\Gamma_L$ based on approximating $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ without stochastic search.

### 3.3.1. A Variational Approximation

Before proceeding, we introduce a bit more notation. For any collection of $L$ particles $\Gamma = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L\}$ and vector $\mathbf{w} = (w_1, \ldots, w_L)$ in the $L$-dimensional simplex, let $q(\cdot \mid \Gamma, \mathbf{w})$ be the discrete distribution that places probability $w_\ell$ on the particle $\boldsymbol{\gamma}_\ell$. Following Ročková

(2018), we will refer to the collection $\Gamma$ as a *particle set* and $\mathbf{w}$ as *importance weights*. Let $\mathcal{Q}_L$ be the collection of all such distributions supported on at most $L$ particles. Finally, for each $\lambda > 0$, let $\Pi_\lambda$ be the tempered marginal posterior with mass function $\pi_\lambda(\boldsymbol{\gamma}) \propto \pi(\boldsymbol{\gamma} \mid \mathbf{y})^{\frac{1}{\lambda}}$. Note that the particles in $\Gamma_L$, which are the $L$ particles with largest posterior mass, are also the $L$ particles with largest tempered posterior mass for all $\lambda$. The following proposition provides the foundation for identifying this collection.

**Proposition 1.** *Suppose that $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ is supported on at least $L$ distinct particles and that $\pi_\lambda(\boldsymbol{\gamma}) \neq \pi_\lambda(\boldsymbol{\gamma}')$ for $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}'$. Let $q^\star_\lambda(\cdot|\Gamma^\star(\lambda), \mathbf{w}^\star(\lambda))$ be the distribution in $\mathcal{Q}_L$ that is closest to $\Pi_\lambda$ in a Kullback-Leibler sense:*

$$q^\star_\lambda = \arg\min_{q \in \mathcal{Q}_L} \left\{ \sum_{\boldsymbol{\gamma}} q(\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\gamma})}{\pi_\lambda(\boldsymbol{\gamma})} \right\}.$$

*Then $\Gamma^\star(\lambda) = \Gamma_L$ and for each $\ell = 1, \ldots, L$, $w^\star_\ell(\lambda) \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{y})^{\frac{1}{\lambda}}$*

*Proof.* See Section B.1 of the appendix. □

In other words, we can find $\Gamma_L$ by finding an approximation of any tempered posterior $\Pi_\lambda$. This is equivalent to solving

$$(\Gamma^\star(\lambda), \mathbf{w}^\star(\lambda)) = \arg\max_{(\Gamma, \mathbf{w})} \left\{ \sum_{\ell=1}^{L} w_\ell \log p(\mathbf{y}, \boldsymbol{\gamma}_\ell) + \lambda H(\Gamma, \mathbf{w}) \right\}, \tag{3.4}$$

where $H(\Gamma, \mathbf{w}) = -\mathbb{E}_q[\log q(\cdot|\Gamma, \mathbf{w})]$ is the entropy of the approximating distribution $q(\cdot|\Gamma, \mathbf{w})$.

Before proceeding, we stress that we are not finding a variational approximation of $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$, the marginal posterior distribution of the continuous parameters of interest. Instead, we approximate the discrete posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$, which places positive probability over all particles $\boldsymbol{\gamma} = (\gamma^\alpha, \gamma^\beta)$, with another discrete distribution $q^\star$ that places positive probability on only $L$ particles.

We pause briefly to reflect on the two terms in Equation (3.4). The first term is, up to

an additive constant depending only on $\mathbf{y}$, the $\mathbf{w}$-weighted average of the height of the log-posterior at each particle in the particle set $\Gamma$. This term is clearly maximized when all of the particles in $\Gamma$ are equal to the MAP. On the other hand, the entropy $H(\Gamma, \mathbf{w})$ of the approximating distribution is maximized when all of the particles in $\Gamma$ are distinct and each $w_\ell = L^{-1}$. The penalty term $\lambda$, which we may also view as an inverse temperature, balances these two opposing forces.

### 3.3.2. Particle Optimization

Finding the global optimum of (3.4) exactly is practically impossible, given the enormous size of the set of all possible particle sets $\Gamma$. Instead, we deploy a coordinate ascent strategy: starting from an initial particle set $\Gamma$ and initial weight vector $\mathbf{w}$, we iteratively update one of $\mathbf{w}$ and $\Gamma$ until we reach a stationary point.

We initialize the particle set by randomly drawing particles $(\hat{\gamma}_K^{(\alpha)}, \hat{\gamma}_{K'}^{(\beta)})$ with replacement where $\hat{\gamma}_K^{(\alpha)}$ is the partition obtained by running k-means on the maximum likelihood estimates of $\boldsymbol{\alpha}$ with $k = K$ clusters. We let $K, K' = 1, \ldots, \lfloor \log(N) \rfloor$. In this initialization, the probability of drawing particle $(\hat{\gamma}_K^{(\alpha)}, \hat{\gamma}_{K'}^{(\beta)})$ is proportional to its marginal posterior probability. Our initialization allows our algorithm to pursue several search directions simultaneously but also allows for some redundancy in the initial particle set. In regions of high posterior probability, such redundancy allows multiple particles to search around a dominant mode, providing a measure of local uncertainty.

Ročková (2018) introduced essentially the same family of optimization problems to identify sparse high-dimensional linear regression models and described a similar coordinate ascent strategy that iteratively updated $\mathbf{w}$ and $\Gamma$. In that work, $\boldsymbol{\gamma}$ was a binary vector indicating which variables to include in the model and the continuous parameters conditional on $\boldsymbol{\gamma}$ were modeled with continuous spike-and-slab priors in the style of George and McCulloch (1993). To update each individual $\boldsymbol{\gamma}_\ell \in \Gamma$, Ročková (2018) restricted attention only to binary vectors which differed in one coordinate. While it is tempting to update each partition in our setting

similarly by re-allocating a single neighborhood to a new or existing cluster, such a strategy is prone to lead to local entrapment.

Indeed, such one-neighborhood updates directly parallel conventional Gibbs samplers for Dirichlet process mixture models (i.e. Algorithms 1 – 8 in Neal (2000)). It is well-known (Celeux et al., 2000) that these samplers can mix very slowly, as their incremental nature make it virtually impossible to pass through regions of low probability between partitions that have similar probability but differ in the cluster assignment of multiple units. In our optimization setting, such a restrictive search strategy results in premature termination at a sub-optimal ensemble $\Gamma$. Instead, a more promising strategy for navigating the space of partitions is to allow multiple elements to be re-allocated at once (Jain and Neal, 2004a). To this end, we consider both *fine* transitions, which re-allocate a single neighborhood to a new or existing cluster (thereby enabling the creation or removal of "islands") and *coarse* transitions, which simultaneously re-allocate multiple neighborhoods.

We have two types of coarse transitions, displayed in Figure 8. The first exchanges multiple neighborhoods simultaneously across a border between adjacent clusters, while the second splits an existing cluster into several sub-clusters and merges some or all of the newly created sub-clusters with other existing clusters. We also consider "merge" moves in which two existing adjacent clusters are combined into a single cluster. These merge moves allow for the removal of islands and the reversal of splits. Sometimes, removing a single neighborhood from a cluster leaves the resulting cluster disconnected. When this happens, we treat the resulting components as individual clusters.

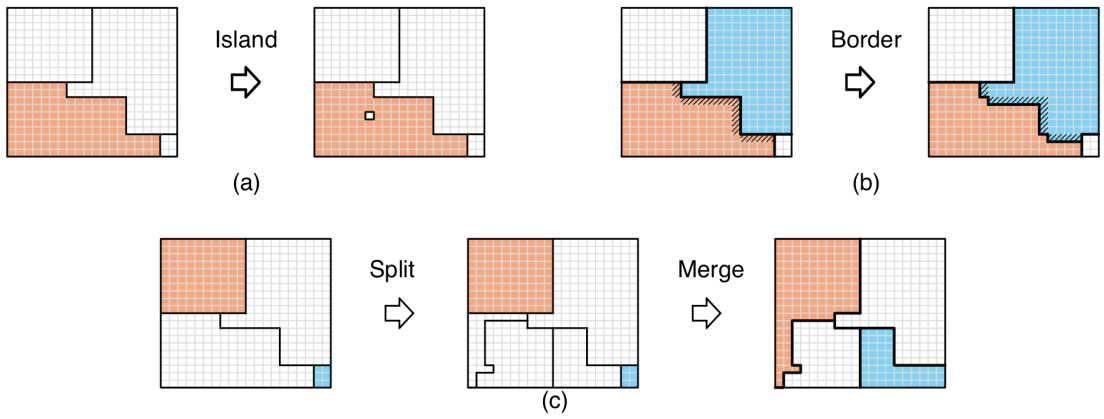


Figure 8: The three broad types of transitions that we consider. An "island" transition (a) removes a single neighborhood from an existing cluster (the lower left orange cluster) and creates a new singleton cluster. A "border" transition moves all neighborhoods at the interface of two adjacent clusters from one cluster to the other. In (b), the neighborhoods moved from the orange cluster to the blue cluster are shaded. The last type of transition (c) first splits an existing cluster (the left cluster in (c)) into multiple parts and then merges some or all of the new sub-clusters into already existing clusters.

In general, we do not attempt all possible coarse and fine transitions while updating a partition. Indeed, there are $O(n)$ possible fine moves and if we allow each of $K$ existing clusters to be split into up to $K_{new}$ sub-clusters, there can be up to $O(K^2 + K \times K_{new}^K)$ possible coarse transitions. Rather than enumerating all of these transitions, we restrict attention to a much smaller set using several heuristics outlined below. For brevity, we describe these heuristics for transitions for $\gamma^{(\alpha)}$; we use exactly the same heuristics for $\gamma^{(\beta)}$.

The conditional conjugacy of our "CAR–within–cluster" model allows us to quickly compute $\mathbb{E}[\alpha_i \mid \boldsymbol{\gamma}, \mathbf{y}]$ and $\mathbb{E}[\overline{\alpha}_k \mid \gamma, \mathbf{y}]$. We use these conditional means as running estimates to propose transitions. For each cluster $k$, we can identify its nearest neighbor $k'$, whose estimated grand cluster mean $\overline{\alpha}_{k'}$ is closest to the estimated grand cluster mean of cluster $k$, $\overline{\alpha}_k$. We then propose exchanging neighborhoods from $k$ across the border between clusters $k$ and $k'$. In this way, we only consider $O(K)$ coarse transitions of the first type. For coarse moves of the second type, which first split an existing cluster into many pieces, we cap the number of new sub-clusters at $K_{new} = 5$. To generate these sub-clusters, we run both k-means and

spectral clustering on the running estimates of the $\alpha_i$'s within the cluster. We also propose splits by removing the top or bottom 5% of these estimates.

Once we split a single cluster into many new sub-clusters, we can identify the nearest neighbor of each sub-cluster among the other existing clusters based on the estimated grand cluster means. We then propose a sequence of merges where a new sub-cluster is merged into its nearest neighbor only if all sub-clusters that are closer to their own nearest neighbors are also merged. For fine transitions, we initially only attempt to remove neighborhood $i$ from its current cluster and move it to a new singleton if its estimated $\alpha_i$ is in the top or bottom 5% of the distribution of estimates within the cluster. Following these heuristics, we consider on the order of $N/10$ fine transitions and $O(K + K \times K_{new}^2)$ total coarse transitions while updating a single partition in our ensemble. During our coordinate ascent algorithm, if we find that none of these transitions are accepted, we then try all $N$ fine moves. This last check ensures that our algorithm converges locally in the sense that no one-tract update to an individual partition will result in a higher objective. While these heuristics are somewhat arbitrary, we have found that they work quite well in practice.

## 3.4. Synthetic Data Evaluation

To investigate the behavior of our proposed optimization procedure, we consider a simpler model of crime $y_{i,t} = \alpha_i + \sigma\varepsilon_{i,t}$ and we place our CAR–within–cluster prior over $\boldsymbol{\alpha}$. We simulate data on a $20 \times 20$ grid of spatial units partitioned into four clusters of sizes 12, 188, 100, and 100. Figure 9 shows the four clusters in the true partition along with three of the different specifications of $\boldsymbol{\alpha}$.

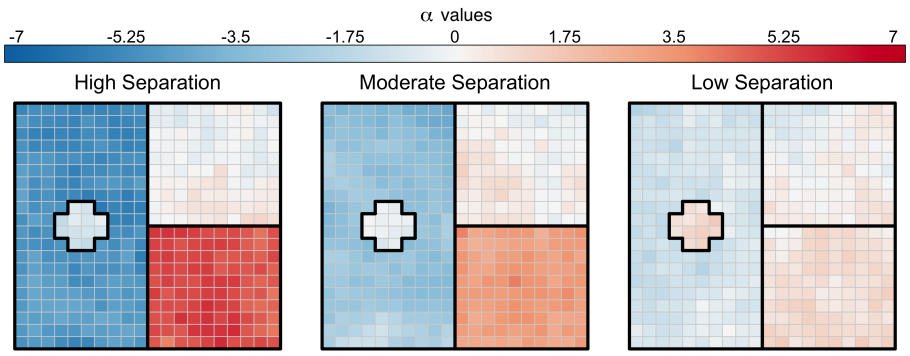


Figure 9: True data generating partition and three different settings of $\boldsymbol{\alpha}$ values. Going from left to right, the distances between the average of the $\alpha_i$'s within each cluster gets progressively smaller. The color of each square corresponds of the true value of $\alpha_i$ used in the synthetic data generating process.

Figure 10 shows the top three partitions recovered when we run our procedure in each of the high, moderate, and low separation settings with two different entropy penalty parameters $\lambda = 1$ and $\lambda = 100$. We placed a truncated Ewens-Pitman prior (3.2) on the latent partition with $\eta = 1$. For this demonstration, we fixed $L = 10, \rho = 0.9$ and set the remaining hyper-parameters according to the heuristics detailed in Section B.2 of the appendix.

It is reassuring to see that when the clusters are well-separated, our method identifies the true partition as the top particle for both values of $\lambda$ and that when the clusters are only moderately separated, the top partitions identified are all quite close to the true partition that generated the data. On the other hand, when there is very little separation between the clusters, the partitions returned by our method are visually quite far from the truth. It turns out that these partitions had substantially more posterior probability than the true partition in this setting.

We know from Proposition 1 that the globally optimal particle set $\Gamma_L^\star$ must (i) contain exactly $L$ particles and (ii) be identical for all values of $\lambda$. We see in Figure 10 that in each of the three settings, the top particles identified for $\lambda = 1$ and $\lambda = 100$ are different. In fact, in the high separation setting, all of the particles in our particle set collapsed to the true partition when $\lambda = 1$. Additionally, in the medium separation setting, the second partition

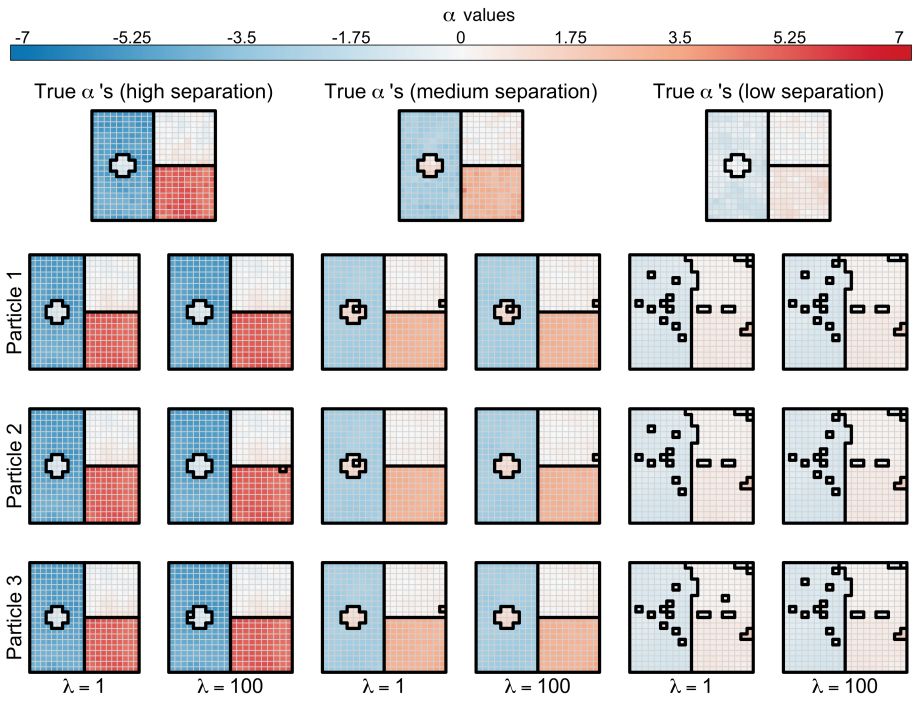


Figure 10: Top three partitions recovered by our particle optimization procedure across different levels of separation of $\boldsymbol{\alpha}$ and values of $\lambda$. The color of each square of the recovered particles corresponds to the value of the posterior mean $\mathbb{E}[\alpha_i \mid \mathbf{y}, \boldsymbol{\gamma}]$. Note, in the high separation setting with $\lambda = 1$, our final particle set contained 10 copies of the same partition.

identified when $\lambda = 1$ is not contained in the particle set obtained when $\lambda = 100$, despite having more posterior probability than all but the top partition in the latter particle set. This behavior, which is at odds with what might be expected from Proposition 1, highlights the local nature of our optimization algorithm.

Recall that the entropy term in Equation (3.4) attempts to offset any potential decrease in posterior probability that accompanies a transition away from a high probability particle set already present in the ensemble to a new particle. The fact that the particle set identified in the high separation setting with $\lambda = 1$ displays extreme redundancy – all of the particles collapsed to the same partition – suggests that this entropy term may not always be sufficient to identify $L$ distinct partitions.

This is not altogether surprising: being bounded from above by $\log L$, the changes in entropy encountered by our algorithm are typically orders of magnitude smaller than changes in the

$\mathbf{w}$-weighted (unnormalized) log-posterior. As we increase $\lambda$ from 1 to 100, however, we find that our procedure recovers $L = 10$ distinct models. In all three settings, we find that some of the particles identified with one choice of $\lambda$ may not be identified with the other choice of $\lambda$, despite having higher posterior probability than many of the particles found with the latter $\lambda$. This could also be an artifact of the local, non-reversible, transitions that we consider. Typically, with larger values of $\lambda$, particles are encouraged to drift to regions of lower posterior probability more forcefully than with lower values of $\lambda$. Moreover, once in those regions, it is typically quite difficult for a particle to "double back" and return to a previously visited state with more posterior probability.

To assess the estimation and partition selection performance of our proposed method quantitatively, we computed the root mean square error (RMSE) of the proposed BMA estimator and the Rand index (Rand, 1971) between the top partition recovered and the true partition averaged over 20 simulated datasets for different choices of cluster separation. The Rand index is defined as the proportion of pairs of elements that are clustered together in both partitions, with values close to one indicating a high degree of similarity between the partitions. Figure 11 shows the average estimation and selection performance for our method run with $\lambda = 1$ along with the following four competitors: (i) the "1-Cluster" model that places all tracts into a single cluster, (ii) the "N-Clusters" model that places all tracts into singleton clusters, (iii) running k-means on the collection of MLE's $\hat{\alpha}_i = \overline{y}_{i,\cdot}$, and (iv) running spectral clustering on these tract averages. When running k-means and spectral clustering, we varied the number of clusters from one to ten. For k-means, we selected the number of clusters using the popular elbow method, and for spectral clustering, we found the number of clusters which minimized the total within-cluster sum of squares. We then computed the conditional posterior expectation $\mathbb{E}[\boldsymbol{\alpha} \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}]$ based on the partition $\hat{\boldsymbol{\gamma}}$ estimated from each of the k-means and spectral clustering procedures. Across our simulations, the estimation and partition selection performance of our method with $\lambda = 100$ was virtually identical to the performance with $\lambda = 1$.

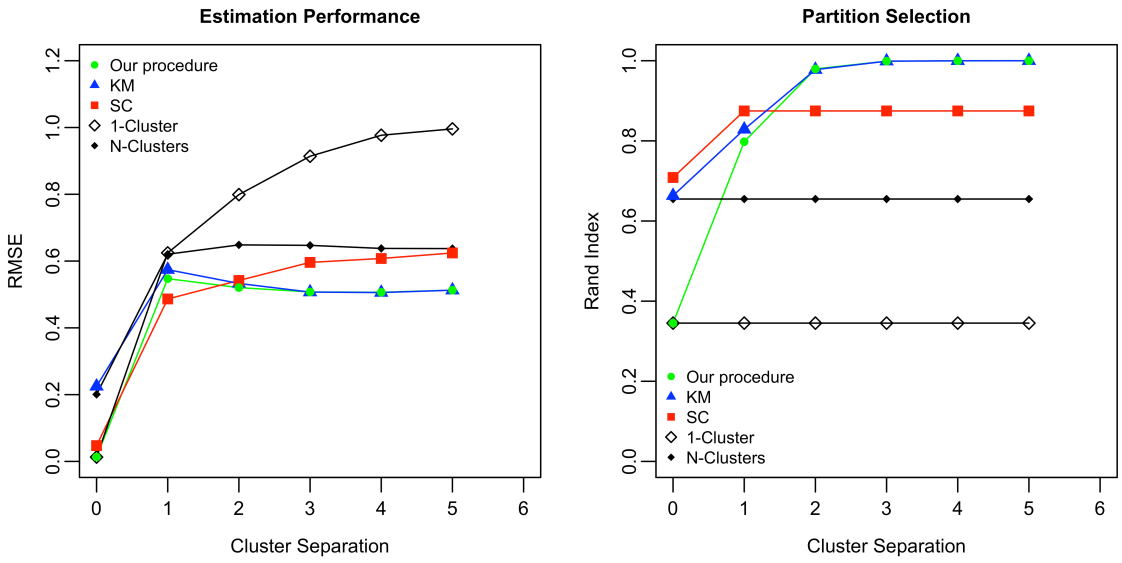


Figure 11: The estimation and partition selection performance, averaged over 20 Monte Carlo simulations, of our method run with $\lambda = 1$ and several competitors across a range of cluster separations.

Immediately we see that, in terms of estimation performance, our procedure is very similar to k-means for non-zero cluster separations. In a certain sense, this behavior is entirely expected when the cluster separation is high: the partition found by k-means in these settings was usually identical to or very close to the true partition, resulting in Rand indices very close to one. However, when the cluster separation is low, our proposed procedure, which identifies several high posterior probability partitions and averages over them, performs much better than k-means, which attempts only to identify a single partition with no reference to the posterior of interest. When there is in fact no separation between the cluster means, the top partition identified by our procedure was always equal to the partition that placed all tracts in a single cluster. In other words, when there truly was no difference between the cluster means, not only does the trivial "1-Cluster" partition have substantially higher posterior probability than other partitions but our particle optimization strategy is also able to recover this partition reliably. This adaptation, in turn, results in excellent estimation performance in this setting.

Interestingly, our method outperforms spectral clustering, in terms of RMSE, except in one setting where the separation between clusters was low but non-zero. In fact, this was the same low separation setting from Figures 9 and 10. As seen in Figure 10, the partitions identified by our method are all quite different than the true partition. It turns out that in this setting, the partition identified by spectral clustering divided the tracts into four equally sized $10 \times 10$ grids; see Figure S1 in Section B.3 of the appendix. This partition is substantially closer to the true partition and it is therefore perhaps not surprising that spectral clustering achieved slightly better RMSE in this setting.

In Figure 11, we showed the RMSE for the full BMA estimator that averaged over all of the particles recovered by our method. Especially when the separation between clusters was very large, often the top partition identified had orders of magnitude more posterior probability than the other partitions identified. This raises a natural question: could we achieve somewhat better estimation performance by averaging over only a subset of the partitions identified by our method instead of averaging over all of them? In our experiments, we found that it was usually better to average over multiple partitions instead of focusing on the MAP plug-in. However, the RMSE was not monotonic in the number of particles averaged over. We also found that the change in RMSE as we varied the number of particles averaged over was quite small, typically of order $10^{-4}$ or less.

## 3.5. Clustering Crime Dynamics in Philadelphia

As described in Section 3.2, we model the transformed number of violent crimes $y_{i,t}$ in neighborhood $i$ at time $t$ as $y_{i,t} = \alpha_i + \beta_i(t - \bar{t}) + \varepsilon_{i,t}$. We further wish to identify two partitions of neighborhoods: one, $\gamma^{(\alpha)}$, that clusters together neighborhoods with similar mean levels of crime $\alpha_i$, and the other, $\gamma^{(\beta)}$, that clusters together neighborhoods with similar time trends $\beta_i$.

For our analysis of the Philadelphia crime data, we consider two priors on the partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$: the Ewen-Pitman prior (3.2) with hyper-parameter $\eta = 5$ and the uniform

prior, both of which are truncated to the set of spatially connected partitions $\mathcal{SP}$. In this analysis, we set $L = 10, \lambda = 100$ and set the remaining hyper-parameters using the heuristics described in Section B.2 of the appendix. The top panel of Figure 12 shows the top three particles recovered when we placed independent Ewens-Pitman priors on each of $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$ while the bottom panel of the figure displays the top three particles recovered with uniform priors on the latent partitions. In Figure 12 we display the top particles as colored maps in which thick lines depict borders between clusters and the color of each neighborhood corresponds to the conditional mean of the $\alpha_i$'s or $\beta_i$'s given the partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$. To illustrate the differences between the identified partitions, we have added greyscale "difference plots" between the colored plots that shade the neighborhoods that are clustered differently. When two partitions are equal, no neighborhoods are shaded in the difference plot.

Similar to our synthetic experiments in Figure 10, some of the identified partitions differ only in the cluster assignment of a small number of neighborhoods. For instance, when we placed a Ewens-Pitman prior on the time-trend partition $\gamma^{(\beta)}$, the top two particles differ in their assignment of a single neighborhood in Northeast Philadelphia. As seen in the second row of Figure 12, while the top particle isolates this neighborhood (labelled A in the figure) in a singleton cluster with a strongly decreasing time trend (ie. large negative $\beta_i$ estimate), the second particle places this neighborhood in a larger cluster with only a moderately decreasing time trend.

Unlike our synthetic experiments, however, some of the identified partitions of the real data differ substantially. This is especially pronounced in the time trend partitions identified when we placed a uniform prior on $\gamma^{(\beta)}$ (bottom row of Figure 12). The first and third particles, for instance, differ substantially in their clustering of neighborhoods in South and West Philadelphia. This difference is most apparent in the large neighborhood (labelled B in the figure) containing the southern sections of the Schuylkill river: the first particle estimates a moderately increasingly time trend in this neighborhood and separates it from

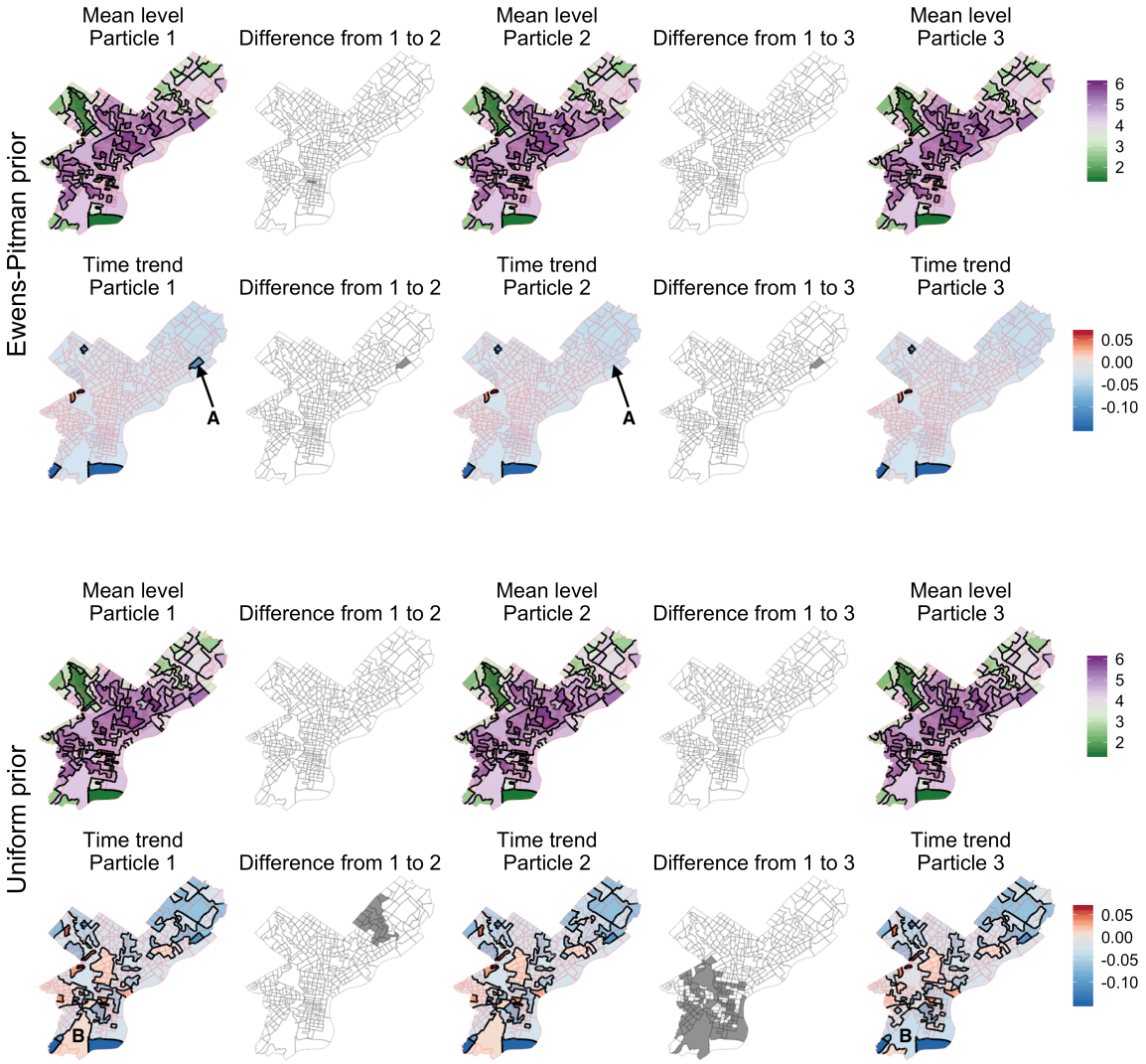


Figure 12: Colored plots: Top three particles identified by our procedure. The thick borders represent the partition, and the color represents the posterior mean of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Black and white plots: difference plots showing in gray the areas where the cluster assignments change between two partitions. **Top:** Ewens-Pitman prior with $\eta = 5$. **Bottom:** Uniform prior on $\mathcal{SP}$.

the neighborhoods immediately to its east and west that have decreasing time trends. In contrast, the third particle clusters all of these neighborhoods together and estimates a decreasing time trend in all of them.

Figure 12 also reveals the sensitivity of the posterior over the partitions to the choice of priors. Specifically, we recover many more clusters when we placed uniform priors on the partitions than when we placed Ewens-Pitman priors. We see this contrast best in the recovered time trend partitions $\gamma^{(\beta)}$. As we see in the second row of Figure 12, with the Ewens-Pitman prior, we recover a relatively small number of clusters: one very large cluster that contains nearly all neighborhoods with an estimated mildly decreasing time trend and a handful of singleton clusters that display more extreme increasing or decreasing time trends. In contrast, with the uniform prior (fourth row of Figure 12), we recover many more clusters. Like with the Ewens-Pitman prior, we still identify some singleton clusters corresponding to more extreme time trends but also identify many moderately sized clusters that display a range of time trends, both increasing and decreasing. Interestingly, though we recover more clusters in the mean level partition $\gamma^{(\alpha)}$ with a uniform prior, the estimates of $\alpha_i$ arising from both priors show little substantive difference.

While Figure 12 compares the top three identified particles, Figure 13 visualizes the overall variation in the entire particle set. Moreover while the former depicts both the parameter estimates in the colored plots and the partition differences in the grayscale plots, the latter only focuses on representing the partition differences. We first depict the top particle and represent the cluster borders with thick lines and then we shade each neighborhood that is assigned to a different cluster in any of the remaining particles $\boldsymbol{\gamma}_{(\ell)}$ for $\ell = 2, \ldots, L$. In this way, we may regard the plots in Figure 13 as the superimposition of the greyscale difference plots computed for each pair $(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(\ell)})$. The left panel of Figure 13 displays the difference in the partitions of the average level of crime $\gamma^{(\alpha)}$ recovered under the Ewens-Pitman prior. From this plot, we see immediately that all of the recovered mean level partitions differ in their cluster assignment of only a small number of neighborhoods. The

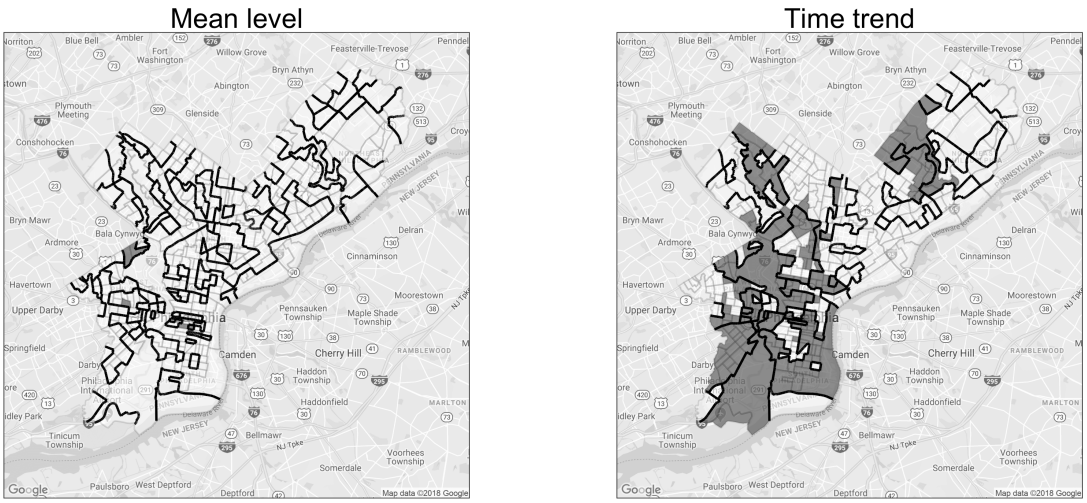


Figure 13: Visualization of the overall partition variation in the particle set. The thick black lines depict the top partition and the shaded areas highlight the neighborhoods that had a different cluster assignment in at least one of the other nine particles. Left panel: variation in $\gamma^{(\alpha)}$ when using the Ewens-Pitman prior. Right panel: variation in $\gamma^{(\beta)}$ when using the uniform prior on $\mathcal{SP}$.

right panel displays a similar representation of the recovered time trend partitions $\gamma^{(\beta)}$ under a uniform prior. We see that there is much more variability in the cluster assignment across the particles, with most of the differences concentrated in South and West Philadelphia. It is not entirely surprising that there is less variability among the partitions recovered using a Ewens-Pitman prior than among partitions recovered with a uniform prior. Essentially, in our local search algorithm, the uniform prior will always favor splitting a large cluster into smaller clusters, even if the corresponding change in marginal likelihood is small. This allows the algorithm to discovery very different partitions with similarly large posterior probabilities. In contrast, the Ewens-Pitman prior tends to favor fine transitions like island moves over coarser moves that simultaneously re-allocate multiple neighborhoods. As a result, the discovered partitions tend to be quite similar to one another under the Ewens-Pitman prior.

Having identified several plausible partitions of the neighborhoods, we now assess the predictive accuracy of our crime model. In addition to the two prior specifications considered in Figure 12, we consider two "hybrid" priors: one in which we place a Ewens-Pitman prior

on $\gamma^{(\alpha)}$ and a uniform prior on $\gamma^{(\beta)}$ and the other in which we place a uniform prior on $\gamma^{(\alpha)}$ and a Ewens-Pitman prior on $\gamma^{(\beta)}$. Figure S2 in appendix B is an analog of Figure 12 that depicts the top three particles identified using these two hybrid priors.

Table 2 reports the out-of-sample RMSE for predicting the level of crime in each neighborhood in 2018 for each of the four different partition prior specifications. The column labelled "Top Particle" reports the RMSE of predictions made using the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from only the top particle (i.e. the MAP estimate of $\boldsymbol{\gamma}$) while the column labelled "BMA" reports the predictions made by averaging over all of the identified particles with BMA. We compare the predictive performance of our method under these four prior specifications to a method that does not impose any shrinkage or clustering and instead makes predictions based only on the maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Table 2: Out-of-sample RMSE using different combinations of priors for the partitions $\gamma^{(\alpha)}$ and $\gamma^{(\beta)}$. The row labelled MLE corresponds to the method which predicts crime in 2018 using the MLE of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ computed using data from 2006 – 2018. The next four rows correspond to the different specifications of the priors on partitions with the prior on $\gamma^{(\alpha)}$ listed first.

| | Top particle | BMA |
|---|---|---|
| MLE | 0.2340 | - |
| EP-EP prior | 0.2568 | 0.2560 |
| Uniform-Uniform prior | 0.2327 | 0.2325 |
| EP-Uniform prior | 0.2339 | 0.2319 |
| Uniform-EP prior | 0.2546 | 0.2539 |

We see that using a uniform prior on $\gamma^{(\beta)}$ yielded better predictive performance than using a Ewens-Pitman prior. As we see in the second row of Figure 12 and the fourth row of Figure S2, with a Ewens-Pitman prior, the vast majority of estimated $\beta_i$'s are negative, indicative of overall average decreasing time trend across the entire city. With a uniform prior (fourth row of Figure 12 and second row of Figure S2), we instead recover a more nuanced picture: while the overall average time trend across the entire city may be negative, there are pockets

of increasing time trends throughout the city. In a certain sense, because the Ewens-Pitman prior strongly discourages the formation of a large number of clusters and instead clusters most of the neighborhoods together, it leads to incorrect estimation of the *sign* of several $\beta_i$'s. Within our simple linear model, incorrect sign estimation can substantially bias future crime forecasts.

In Table 2, we see that placing a Ewens-Pitman prior on $\gamma^{(\alpha)}$ and a uniform prior on $\gamma^{(\beta)}$ yielded the best predictive performance. Figure 14 depicts the top particle identified under this prior specification.

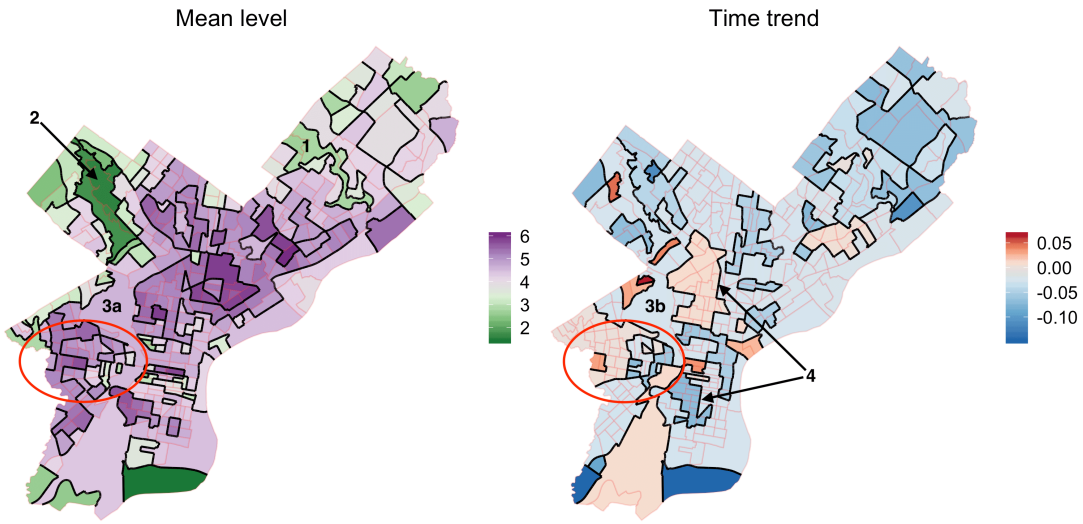


Figure 14: Partitions (thick black lines) and posterior mean coefficients (colors) in the top particle under the EP prior on $\gamma^{(\alpha)}$ and uniform prior on $\gamma^{(\beta)}$, which is the model with the smallest prediction error (in the BMA sense).

We recognize many aspects of Philadelphia's geography directly from the cluster structure shown in Figure 14. For instance, the clusters labelled **1** and **2** in the figure correspond to the areas surrounding the Pennypack and Wissahickon rivers, respectively. Further, several cluster borders coincide exactly with the boundaries of Fairmount Park (indicated by **3a** and **3b** in the figure) and the major arterial road Broad Street (labelled **4**).

We finally analyze the West Philadelphia and University City region (circled in Figure 14),

which contains both Drexel University and the University of Pennsylvania. For the most part, this region is characterized by relatively high levels of crime (darker shades of purple in the left panel of the figure) with the exception of two neighborhoods that are immediately adjacent to the universities (lighter green shades). There is substantial heterogeneity in the estimated time trend within the region as well. Notably, we estimate a decreasing trend in the neighborhoods in the immediate vicinity of the universities and a slightly increasing trend further away from the universities. This finding aligns with previous reports of the positive impact of the University of Pennsylvania's West Philadelphia Initiatives aimed at improving the social and economic landscape around the university campus (Ehlenz, 2016).

## 3.6. Discussion

Accurate estimation of the change in crime over time is a critical first step towards a better understanding of public safety in large urban environments. An especially important challenge to such estimation is the potential presence of sharp discontinuities, which may be smoothed over by naive spatial shrinkage procedures. Focusing on the city of Philadelphia, we introduced a Bayesian hierarchical model that naturally identifies these discontinuities by partitioning the city into several clusters of neighborhoods and introduces spatial smoothness within but not between clusters. In particular, we focused on recovering two latent spatial partitions, one for the mean-level of crime over the twelve year period 2006 – 2017 and one for the time-trend.

Rather than use a computationally prohibitive stochastic search, we instead sought to identify partitions with highest posterior probability by solving a single optimization problem. We showed that optimizing the proposed objective function is formally equivalent to finding a particular variational objective and introduced a local search strategy for solving this problem. While our primary focus has been on crime in the city of Philadelphia, our ensemble optimization framework is more general and there are a number of areas of future development, which we discuss below.

The results of our applied analysis were quite sensitive to the choice of prior placed on the underlying spatial partition. With a Ewens-Pitman prior, nearly all of the neighborhood time trends were assigned to a single cluster while with a uniform prior, we obtained a much richer cluster structure. It would be interesting to construct an objective prior for spatial partitions along the lines of Casella et al. (2004).

While it may be sufficient to consider a linear temporal model of crime when there are relatively few time points (Bernardinelli et al., 1995; Anderson et al., 2017), with more observations per census tract, it is reasonable to consider more flexible models. For instance, we could model $y_{i,t} \sim N(f_i(\mathbf{x}_{i,t}), \sigma^2)$ and place Gaussian process priors over the $f_i$'s within each cluster. Such an elaboration retains conditional conjugacy and we can still use our ensemble optimization strategy to identify clusters with high posterior probability, though computing the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ is somewhat more involved. It is more difficult to deploy our ensemble optimization strategy directly when the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ is not available in closed-form. While it may sometimes be possible to use an EM algorithm like Ročková (2018), this is not always feasible for more complicated models. One very natural idea would be to estimate the marginal likelihood with a Laplace approximation.

In many of our empirical examples and especially when we used Ewens-Pitman priors, the particle set can remain stuck in the vicinity of a dominant mode. When this happen, it is not immediately obvious whether the posterior truly concentrates around a single dominant mode or if there are other pockets of substantial posterior mass that are far away. Unfortunately, the entropy term in Equation 3.4 may provide insufficient repulsion between the particles to probe this latter possibility. Operationally, the entropy term discourages redundancy in the particle set by penalizing exact equality between particles but does not penalize placing a particle in the vicinity of another model that is already present in the particle set. One way around this potential weakness is to augment the optimization objective in (3.4) with an additional penalty term that directly penalizes the pairwise distance between particles in the particle set. In doing so, however, we would lose the guarantee of

optimality afforded by Proposition 1.

# Chapter 4

# Prior Choice for Clustering Areal Data

## 4.1. Introduction

Areal data, a kind of spatial data that are measured at an aggregate level within regions, are common in many disciplines, from socio-economic census data and health data, which are aggregated for privacy reasons, to image data, where light and color are collected at the pixel level. These regions, which partition the space of interest and have well defined borders, are called *areal units*.

It is often of interest to find groups of areal units that display homogeneous characteristics, to summarize the spatial variation and gain insight in underlying patters, or for better estimation through the use of mixture models. These groups, or *clusters*, of areal units are often presumed to be formed by adjacent units, or to be geographically *connected*, as near units are expected to have more similar behaviors than distant ones.

A fairly well known application of spatial clustering of areal data is disease mapping. The number of disease events are aggregated in each geographical region (e.g. counties) for privacy concerns and the disease risk is estimated in each areas, in order to map patterns and identify parts of high/low risk. For a review of disease mapping see for example Best et al. (2005); Lawson (2013). Another interesting and similar application is the study of crime change over time, as we describe in Chapter 3.

In all of these applications, the interest lies in the detection of clusters of regions that exhibit markedly different trends, or equivalently, of the underlying partition of areal units. The Bayesian perspective is especially helpful in this context, because it allows the expression of uncertainty and of prior information about this latent partition through a prior distribution, or model.

Several approaches have been considered in the literature for modeling this partition. Approaches based on tessellation techniques (Knorr-Held and Raßer, 2000; Denison and Holmes, 2001; Feng et al., 2016) and on the Potts model (Green and Richardson, 2002; François et al., 2006; Johnson et al., 2013), as well as other approaches (Anderson et al., 2017) first choose a fix number of clusters or treat it as a random variable, and conditionally model the cluster membership of each unit.

In contrast to these approaches, it is possible to consider the latent clustering as an object living in the space of partitions. This approach requires the use of distributions over the space of partitions. A well-known choice is the distribution induced by the Dirichlet Process (Ferguson, 1973), also known as the Ewens distribution (Ewens, 1972; Pitman, 1996). Among the reasons for its widespread use in the Bayesian Nonparametric community, is its mathematical convenience; specifically, under the Ewens prior, the partitions are *exchangeable*, i.e. the probability does not depend on the label of the units.

However, spatial data are not exchangeable: as suggested by Tobler's first law of geography, "everything is related to everything else, but near things are more related than distant things"; thus, the probability that two adjacent units belong to the same cluster should be higher, not the same, as the one for two units that are distant, because *a priori* we expect the two near units to be more similar. In other words, a distribution that is invariant to relabeling of the units ignores the spatial structure. For this precise reason, a prior distribution that induces exchangeable partitions is not appropriate for clustering areal data. Instead, several alternatives have been proposed, which generate *spatial partitions*, in which adjacent units have a larger probability of being clustered together.

The majority of these approaches start from the Dirichlet Process and the distribution it induces on partitions, and alter it to account for the spatial information. This can be done in two ways: the Dirichlet Process can itself be changed, by modifying the *stick-breaking* construction (see e.g. Ren et al., 2011; Jo et al., 2017; Hosseinpouri and Khaledi, 2019), or by directly modifying the distribution induced on partitions (see e.g. Orbanz and Buhmann,

2008; Dahl et al., 2017; Page et al., 2016). We will focus on the latter approach, as it is very difficult to study the distribution induced on partitions from the former approach.

Nonetheless, the non-exchangeability of spatial partitions makes them less mathematically convenient, and thus the properties of many of these distributions have not been studied thoroughly. However, for a practitioner it is fundamental to know which prior properties each of them encode and thus which distribution is more appropriate for a particular application or problem.

One important mathematical property is the property of coherence across sample sizes. Unfortunately, often distributions over non-exchangeable partitions lack this property. A notable exception can be found in the PPMx distribution of Müller et al. (2011), which can be proven to satisfy this property. The PPMx introduces a general framework to adapt distributions over partitions, in particular Product Partition Models (PPM), in the presence of covariates, and can be adapted to several settings. In particular, it has has been extended for spatial partitions, but only in the case of point-referenced data (Page et al., 2016). The case of areal data has not been considered.

In this work, we have a dual goal: we explore the behavior induced by different choices of these prior distributions, and we compare their properties; moreover, we extend the work of Page et al. (2016) to define Spatial Product Partition Models for areal data. We introduce two distributions for partitions of areal data. We show that the first satisfies the property of coherence across sample sizes, while the second does not but displays more favorable empirical properties.

The rest of this work is organized as follows. In section 4.2 we review some of the most popular distributions over partitions, from exchangeable partitions, to partitions of data with covariates, to partitions of spatial areal data. In section 4.3 we describe the two new distributions over partitions, designed to extend the PPMx framework to the case of areal data. Lastly in sections 4.4 and 4.5 we analyze some prior and posterior properties of the

distributions considered.

## 4.2. Background

### 4.2.1. Priors for exchangeable partitions

The Ewens distribution (Ewens, 1972; Pitman, 1995, 1996), also known as the distribution over partitions induced by the Dirichlet Process (DP) (Ferguson, 1973), is defined by

$$\Pr(\rho = \{S_1, \ldots, S_k\}) = p(n_1, \ldots, n_k) = \frac{\alpha^k \prod_{j=1}^{k}(n_j - 1)!}{\alpha(\alpha+1)\ldots(\alpha+n-1)} \tag{4.1}$$

where $\rho$ represents a partition of $n$ elements, $S_j$ represents one of its part, or *cluster*, and $n_j$ the cardinality of $S_j$, $n_j = |S_j|$. This formula shows that under this distribution the probability of a partition depends only on the total number of elements $n$, on the number of clusters $k$ and their sizes $n_j$. As a consequence, the partitions generated by the Ewens distribution are exchangeable and the formula 4.1 is called exchangeable partition probability function (EPPF) (Pitman et al., 2002). The distribution additionally depends on the value of the parameter $\alpha$, which influences the number of clusters generated; as $n \to +\infty$, the number of clusters grows as $\alpha \log(n)$, where $n$ is the number of elements that are being partitioned.

The Ewens distribution is very popular for its mathematically simplicity; however, this logarithmic growing rate often does not match the behavior of real world phenomena. We can consider an extension of the Ewens distribution, the Ewens-Pitman[1] distribution (Pitman and Yor, 1997), which is also known as the distribution induced by the Pitman-Yor Process. The Ewens-Pitman distribution over partitions is defined by

$$\Pr(\rho = \{S_1, \ldots, S_k\}) = \frac{\prod_{j=1}^{k-1}(\alpha + j\sigma)}{(\alpha+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}.$$

[1]There is some ambiguity in the literature about the names for these distributions. Casella et al. (2014) refers to the distribution induced by the Dirichlet Process as Ewens-Pitman distribution. Here we follow what seems to be the most common choice, following also the work of Dahl et al. (2017).

where $(x)_m = x(x+1)\dots(x+m-1) = \Gamma(x+m)/\Gamma(x)$. The additional parameter $\sigma \in [0,1)$ is affecting the asymptotic number of clusters: it now grows as $n^\sigma$. This *power law* behavior is considered more appropriate for many real world applications. Similarly to the Ewens distribution, the Ewens-Pitman probability function only depends on the clusters sizes and it induces exchangeable partitions.

Both the Ewens and the Ewens-Pitman prior belong to the family of *Product Partition Models* (PPM), introduced by Hartigan (1990); Barry and Hartigan (1992). These distributions are characterized by a probability function that factorizes over the clusters:

$$\Pr(\rho = \{S_1, \dots, S_k\}) = \frac{1}{K} \prod_{j=1}^{k} c(S_j),$$

where $c(S)$ is a cohesion for each cluster $S \in \rho$ and $K$ is the normalizing constant

### 4.2.2. Priors for partitions dependent on covariates

Exchangeable partitions are mathematically convenient, but cannot be used to describe certain situations. For example, in the presence of covariates it is not advisable to assume the probability to be invariant to permutations of the units indices. In similar works, Müller et al. (2011) and Park and Dunson (2010), extend the Product Partition Model (PPM) to allow for predictor-dependent partitions, and they introduce analogous models: the PPM with covariates (PPMx) and the generalized product partition model (GPPM).

They consider a setting in which each subject's response $y_i$ is observed together with predictors $x_i$; their objective is to model the responses given a partition of the elements and to provide a prior distribution for the partition which accounts for the covariates. Let $\mathbf{x}^n = (x_1, \dots, x_n)$ and $\mathbf{y}^n = (y_1, \dots, y_n)$ denote the predictors and responses for all the units under consideration; moreover let $\mathbf{x}_S = (x_i : i \in S)$ and $\mathbf{y}_S = (y_i : i \in S)$ be the collection of predictors and responses for the units in cluster $S$. The model for the responses is a mixture model; given a partition $\rho$ of the units, it treats the data in different clusters

as conditionally independent, i.e. the likelihood factorizes over the clusters:

$$p(\mathbf{y}^n|\rho = \{S_1, \ldots, S_k\}, \boldsymbol{\theta}) = \prod_{j=1}^{k} p(\mathbf{y}_{S_j}|\theta_j).$$

The prior for the partitions is constructed as a product over the clusters of a term that operates on the cluster itself $c(S_j)$ and of a predictor-dependent cohesion function $g(\mathbf{x}_{S_j})$ that operates on the covariates of the units in that cluster:

$$p(\rho = \{S_1, \ldots, S_k\}|\mathbf{x}^n) \propto \prod_{j=1}^{k} c(S_j)g(\mathbf{x}_{S_j}).$$

While the term $c(\cdot)$ was referred to as "cohesion" under the Product Partition Model, to avoid any ambiguity from now on we will use the term cohesion only to refer to the predictor-dependent cohesion function $g$. The first term $c(\cdot)$ can be borrowed from the most popular PPM's, such as the Ewens or the Ewens-Pitman distribution. There are instead several ways of constructing the cohesion functions $g(\cdot)$ for the covariates. The definition proposed by Müller et al. (2011) and Park and Dunson (2010) uses an auxiliary model for the covariates given their cluster membership: assume that for each cluster the covariates are exchangeable and generated from the auxiliary model $q(\cdot|\xi)$; then set $g(\mathbf{x}_{S_j})$ as the marginal likelihood for the covariates:

$$g(\mathbf{x}_{S_j}) = \int \prod_{i \in S_j} q(x_i|\xi_j)q(\xi_j)d\xi_j. \tag{4.2}$$

By assuming that in the auxiliary model the covariates are exchangeable within each cluster, the prior distribution will favor partitions whose clusters have similar values of the covariates.

Note that the distribution for partitions defined by the PPMx and GPPM could be interpreted as an application of Bayes' formula:

$$p(\rho|\mathbf{x}^n) \propto p(\mathbf{x}^n|\rho)p(\rho)$$

where $p(\rho) \propto \prod_{j=1}^{k} c(S_j)$ is given by a distribution for exchangeable partitions, such as the Ewens distribution, and $p(\mathbf{x}^n|\rho) = \prod_{j=1}^{k} \int \prod_{i \in S_j} q(x_i|\xi_j) q(\xi_j) d\xi_j$ is the auxiliary mixture model for the covariates. Considering the *posterior* distribution $p(\rho|\mathbf{x}^n)$ of this auxiliary model, leads to the construction of a distribution for non-exchangeable partitions.

One of the important properties of the PPMx is its coherence across sample sizes, which relates the probabilities of the partition over $n$ elements $\rho_n$ and the partition over $n+1$ elements $\rho_{n+1}$:

$$p(\rho_n|\mathbf{x}^n) = \sum_{z_{n+1}} \int p(\rho_{n+1}|\mathbf{x}^{n+1}) q(x_{n+1}|\mathbf{x}^n) dx_{n+1} \tag{4.3}$$

where $q(x_{n+1}|\mathbf{x}^n) \propto g(\mathbf{x}^{n+1})/g(\mathbf{x}^n)$.

The generality of this formulation makes this model applicable in many applications. In fact, the PPMx prior has been adapted to many different kinds of covariates, and in particular to spatial data: Page et al. (2016) specialize this model to the case of point-referenced data, i.e. data representing a point that can vary continuously in space, such as GPS locations. However, they do not consider the case of areal data. In the next section we describe the existing approaches to model partitions of spatial data, both point-referenced and areal.

### 4.2.3. Prior for partitions of spatial data

Spatial data is a particular case that falls into the framework of Müller et al. (2011) and Park and Dunson (2010), where the covariate $\mathbf{x}$ represent the spatial location of the observations. Because of the ubiquity of such data, specific models have been constructed to deal with data.

For example, Page et al. (2016) adapt the PPMx framework to point-referenced spatial data, i.e. the covariates $\mathbf{x}$ represent the GPS location of the observations. They construct the auxiliary model by assuming a Normal-Inverse-Gamma prior for the location $x_i$. Thus, if $\xi_j = (\mu_j, \Sigma_j)$, then $q(x_i|\xi_j) = N(\mu_j, \Sigma_j)$ and $q(\xi_j) = NIG(\mu_j, \Sigma_j)$.

While this auxiliary model could artificially be used on areal data by considering the coor-

dinates of areal units' centroids, it is not appropriate for this kind of data. As an example, consider neighborhoods in a city, which form a non-regular graph and where some neighborhoods can be much larger than the others; specifically, if one large neighborhood is surrounded by many small neighborhoods, the distance between the centroids of the large and a small neighborhood would be much larger than the distance between the centroids of two small adjacent neighborhoods.

It is thus important to consider auxiliary models that specifically address the case of areal data; however, Page et al. (2016) do not propose auxiliary models for this case. Instead, they provide alternative distributions, some of these obtained by combining a cohesion function with a standard distribution. However, these alternative distribution don't have the property of coherence of the PPMx. We now describe a notable distribution, the Ewens or DP distribution restricted on connected partitions, which was also described in Page et al. (2016). Consider the definition of *connected cluster* and of *connected partition*, which we adapt from Page et al. (2016).

**Definition 1** (Connected cluster)**.** Let $d(\cdot, \cdot)$ be the distance between areal units given by the length of the shortest path connecting them. Consider cluster $S_h \in \rho$. We say that cluster $S_h$ is spatially connected if there does not exist $i \in S_h$ and $i' \notin S_h$ such that for all $j \in S_h$, $d(i, i') < d(i, j)$.

Note that if a cluster is *not connected*, then there exist some $i, i'$ such that $d(i, j) > d(i, i') \geq 1$, for all $j \in S_h$. In other words, if a cluster is not connected under this metric, there must exist $i \in S_h$ such that $d(i, j) > 1$ for all $j \in S_h$, i.e. the shortest path between $i$ and all the other elements of the clusters has at least length 2, meaning that *there is no path fully contained in the cluster connecting its elements*. Moreover, this definition depends on both the clusters and the information about the adjacency structure, which is encoded in the adjacency matrix $W$.

**Definition 2** (Connected partition)**.** A partition is said spatially connected if all of its clusters are spatially connected.

With these definitions, we can consider a truncation function $t(\rho, W) \propto 1$ if $\rho$ is connected and 0 otherwise, and restrict any distribution over partitions to the subset of spatially connected partitions. In particular, we can consider the Ewens or DP distribution and only assign positive probability to connected partitions:

$$\Pr(\rho = \{S_1, \ldots, S_k\}|W) \propto p_{DP}(\rho)t(\rho, W) \propto \begin{cases} \alpha^k \prod_{j=1}^{k}(n_j - 1)! & \text{if } \rho \text{ connected} \\ 0 & \text{otherwise.} \end{cases}$$

We will refer to this distribution as Connected-DP. However, as citepage2016spatial point out, this definition is "intuitively appealing" but "challenging to implement from a computational stand point".

Another distribution based on a cohesion function is the Markov Random Field DP (Orbanz and Buhmann, 2008). Under this distribution, the cohesion function is constructed to satisfy the Markov property, using Hammersley-Clifford theorem. Specifically:

$$p(\rho = \{S_1, \ldots, S_k\}|W) \propto \prod_{h=1}^{k} c(S_h) e^{-\lambda \sum_{i,j \in S_h} w_{ij}}. \tag{4.4}$$

If we derive the conditional distribution for the cluster membership $z_i$, we see that the cohesion term only depends on the cluster membership of its neighbors, i.e. the units $j$ for which $w_{ij} = 1$: $p(z_i|\mathbf{z}_{-i}, W) \propto c(z_i|\mathbf{z}_{-i}) \exp(-\lambda \sum_j w_{ij}\delta_{z_i = z_j})$.

All of these distributions, as the majority of the distributions for non-exchangeable partitions, are unfortunately specified up to a normalizing constant, which is analytically intractable and cannot be computed. A different approach instead is taken by Dahl et al. (2017), who specifies the distribution as the product of conditional probabilities and thus can compute it exactly. However, since the distribution induced is non-exchangeable, the probability depends on the order of the units. In other words, it depends on a permutation $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ of the units $\{1, \ldots, n\}$. Given a pairwise similarity function, $\lambda(i, j)$, that

measures the vicinity of two units, and given a permutation $\boldsymbol{\sigma}$, Dahl et al. (2017) define[2] the Ewens-Pitman Attraction distribution (EPA) as

$$p(\rho = \{S_1, \dots, S_k\}|\lambda, \boldsymbol{\sigma}) = \prod_{i=1}^{n} p(z_{\sigma_i}|z_{\sigma_1}, \dots z_{\sigma_{i-1}}, \lambda, \boldsymbol{\sigma})$$

where

$$p(z_{\sigma_i} = h|z_{\sigma_1}, \dots z_{\sigma_{i-1}}, \lambda, \boldsymbol{\sigma}) = \begin{cases} \frac{i-1}{\alpha+i-1} \frac{\sum_{j<i} \lambda(\sigma_i,\sigma_j)\mathbf{1}(z_{\sigma_j}=h)}{\sum_{j<i} \lambda(\sigma_i,\sigma_j)} & \text{if } h \in \{z_{\sigma_1}, \dots, z_{\sigma_{i-1}}\} \\ \frac{\alpha}{\alpha+i-1} & \text{if } h \text{ new cluster.} \end{cases} \tag{4.5}$$

This clever construction preserves the asymptotic distribution of the total number of clusters. In fact, the probability of creating a new cluster given the assignment of the first $i-1$ units is the same as the one induced by the Ewens distribution: $\alpha/(\alpha+i-1)$. At the same time, the probability of joining a previous cluster is weighted by the similarity of the units belonging to it. In the simple case that $\lambda$ is constant, the EPA reduces to the conditional specification of the Ewens distribution: $p(z_{\sigma_i} = h|z_{\sigma_1}, \dots z_{\sigma_{i-1}}, \boldsymbol{\sigma}) = n_i/(\alpha+i-1)$. The EPA distribution can easily be used in areal data settings, where the similarity function can be constructed as a non-increasing function of the distance $d$, for example $\lambda(d) = \exp(-\tau d)$. In the case of areal data, we can consider the distance defined by length of the shortest path connecting two areal units.

To avoid the definition of the distribution depend on the permutation of the units, $\boldsymbol{\sigma}$ is assumed to be uniformly distributed and the marginal distribution on partitions can be computed as

$$p(\rho = \{S_1, \dots, S_k\}|\lambda) \propto \sum_{\boldsymbol{\sigma}} p(\rho = \{S_1, \dots, S_k\}|\lambda, \boldsymbol{\sigma}).$$

Thus, when the number of units is not small, the functional form cannot be computed, as it depends on an intractable sum.

[2]The definition given in Dahl et al. (2017) is more general and includes an additional parameter $\delta$, equivalent to the additional parameter present in the Ewens-Pitman distribution; here we report the particular case where $\delta = 0$.

## 4.3. Our proposal: priors for partitions of areal data

Differently from point-referenced data, which are provided with GPS locations or continuous coordinate on a space, areal data consists of data measured or aggregated within a discrete set of regions, called *areal units*. Their spatial information is described by their adjacency relationship, which defines the neighborhood structure. This information is often encoded in the adjacency matrix $W$, whose entries $w_{ij}$ are equal to 1 when units $i$ and $j$ share a border, and 0 otherwise.

Because of the adjacency structure, areal data can be viewed as a graph: each areal unit can be interpreted as a node, and nodes corresponding to neighboring units can be thought as connected by an edge.

Moreover, information concerns pairs of units, since it is encoded in the pairwise adjacency relationship. As a consequence, it is impossible to create a model that factorizes over clusters, unless the pairwise information between different clusters is ignored or counted twice. Since both options do not seem advisable, we argue that the Product Partition Model assumption is not appropriate for partitions of areal data.

Instead, we suggest to consider a cohesion function that does not factorize over the clusters $g(\mathbf{x}^n)$:

$$p(\{S_1, \ldots, S_k\}|\mathbf{x}^n) \propto g(\mathbf{x}^n) \prod_{j=1}^{k} c(S_k).$$

This distribution is consistent with the idea of using Bayes rule to define the prior for partitions, $p(\rho|\mathbf{x}^n) \propto p(\mathbf{x}^n|\rho)p(\rho)$, with the difference that in this case the auxiliary model for the covariates $p(\mathbf{x}^n|\rho)$ does not factorize over the clusters.

To construct $g$ we first need to decide what is the covariate information $\mathbf{x}^n$ that should be used to construct the distribution; two options are described in the following paragraphs: we could use the adjacency matrix $W$, or we could use an appropriate summary statistic.

**SBM-PPMx** Given a collection of $n$ areal units, we want to model their adjacency structure by assigning a probability distribution to the entries of $W$, which represent the relationship of being adjacent to another unit. In other words, we want to construct an auxiliary model for $\mathbf{x}^n = W$. Since the adjacency matrix $W$ is symmetric, only half of the entries need to be modeled: $\{w_{ij} : i < j\}$. With an abuse of notation this upper triangular matrix will be denoted $W$. Moreover, to emphasize the dependence on the sample size, we will sometimes use the notation $W^n$.

We will borrow the terminology from graph theory, and say that $w_{ij} = 1$ signifies the presence of an *edge* between unit $i$ and unit $j$. To model the distribution of $W$, we can assign an ordering to the units and work iteratively: consider the edges between the"last" unit $n$ and the previous units $1, 2, \ldots, n-1$: $\mathbf{w}_n = \{w_{i,n} : i < n\}$. If we denote with $W^n$ the adjacency matrix of all the $n$ units, and with $W^{n-1}$ the matrix for the first $n-1$ units, we can define $p(W^n|\rho, \theta) = p(\mathbf{w}_n|\rho, W^{n-1}, \theta)p(W^{n-1}|\rho, \theta)$. To make this more intuitive we can think of describing the adjacency structure of the neighborhoods of a city by starting from the center and adding one region at a time, working in circles moving further away from the center. Every time we consider an additional unit, we examine the adjacency relationship to the previous units.

Let us now define the distribution of $p(\mathbf{w}_n|\rho, W^{n-1}, \theta)$. Remember that $w_{in} = 1$ if units $i$ and $n$ share a border and 0 otherwise; moreover, let $\rho$ be the partition of the units and let $z_1, \ldots, z_n$ represent the cluster memberships. To define our model, we borrow the idea underlying the Stochastic block model (SBM) (Holland et al., 1983): nodes in the same group are more connected than nodes in different groups. Thus, conditional on the partition of the units, we can expect a larger number of edges between units in the same cluster, and a smaller number of connections between different clusters. Thus we can divide the units $1, 2, \ldots, n-1$ into two groups: the units in the same cluster as unit $n$ and those in different clusters; let $s_n$ be the cardinality of the first group, $s_n = \#\{i : z_i = z_n\}$. As in the SBM, we assume the elements of $\mathbf{w}_n$ to be independent and we model each of them as

Bernoulli random variables with probability $\xi_1$ if they belong in the same cluster as $n$, and with probability $\xi_0$ if they are in a different cluster. The parameter of this auxiliary model then are $\theta = (\xi_1, \xi_0)$. Thus,

$$p(\{w_{in} : z_i = z_n\}|\rho, W^{n-1}, \theta) = \prod_{i:z_i=z_n} \xi_1^{w_{in}}(1-\xi_1)^{1-w_{in}}$$
$$p(\{w_{in} : z_i \neq z_n\}|\rho, W^{n-1}, \theta) = \prod_{i:z_i\neq z_n} \xi_0^{w_{in}}(1-\xi_0)^{1-w_{in}}$$

and letting $v_n = \sum_{z_i=z_n} w_{in}$ and $d_n = \sum_i w_{in}$ be respectively the within-cluster neighbors and the total number of neighbors (the *degree*) of $n$,

$$p(\mathbf{w}_n|\rho, W^{n-1}, \theta) = \xi_1^{v_n}(1-\xi_1)^{s_n-v_n} \cdot \xi_0^{e_n-v_n}(1-\xi)^{n-1-s_n-(e_n-v_n)}.$$

If we complete this model iteratively, we find that the entries of $W^n$ can be divided in two groups: the within cluster connections $\{w_{ij} : z_i = z_j, i < j\}$, and the between clusters connections $\{w_{ij} : z_i \neq z_j, i < j\}$. The distribution of the whole matrix is then

$$p(W|\rho_n, \theta) = \prod_{i<j:z_i=z_j} \xi_1^{w_{ij}}(1-\xi_1)^{1-w_{ij}} \prod_{i<j:z_i\neq z_j} \xi_0^{w_{ij}}(1-\xi_0)^{1-w_{ij}} =$$
$$= \xi_1^{v^n}(1-\xi_1)^{s^n-v^n} \cdot \xi_0^{e^n-v^n}(1-\xi)^{n(n-1)/2-s^n-(e^n-v^n)},$$

where $s^n = \sum_{i\leq n} s_n$, $v^n = \sum_{i\leq n} v_n$ and $e^n = \sum_{i\leq n} e_n$. Note that despite we describe the construction in an iterative way, the auxiliary does not depend on the order of the units.

To complete our auxiliary model, we specify the prior for $\theta$: $p(\xi_1) = \text{Beta}(a, b)$ and $p(\xi_1) = \text{Beta}(c, d)$. In Appendix C.1 we give a heuristic to specify the hyper-parameters so that $\xi_0 < \xi_1$.

Finally, we can derive the functional form of the cohesion function under the SBM auxiliary

model:

$$g(\mathbf{x}^n) = g(W^n) = \frac{B(a+v^n, b+s^n)}{B(a,b)} \frac{B(c+e^n-v^n, d+n(n-1)/2-s^n)}{B(c,d)}. \tag{4.6}$$

Since we are able to consider a variable for each unit, i.e. $\mathbf{w}_n = \{w_{i,n} : i < n\}$, the coherence across sample sizes described in equation (4.3) still sholds.

**Proposition 1.** *The distribution over partitions induced by the SBM-PPMx auxiliary model is coherent across sample sizes, in the sense of*

$$p(\rho_n|W^n) = \sum_{z_{n+1}} \int p(\rho_{n+1}|W^{n+1}) q(\mathbf{w}_{n+1}|W^n) d\mathbf{w}_{n+1}$$

*where* $q(\mathbf{w}_{n+1}|W^n) \propto g(W^{n+1})/g(W^n)$.

**Areal-PPMx** The SBM auxiliary model introduced above treats the entries of the matrix as independent; this assumption might be inappropriate and lead to undesirable properties in the cohesion function induced. A simple way to introduce dependence is to model the data through a summary statistic; a natural choice in this setting is given by the number of within cluster connections $v^n = \frac{1}{2}\sum_{ij} w_{ij}\mathbf{1}\{z_i = z_j\}$. Since $v^n$ is bounded by the total number of edges $e^n = \frac{1}{2}\sum_{ij} w_{ij}$, we can easily model $v^n$ as a Binomial distribution with probability $\theta$:

$$p(v^n|e^n, \rho, \theta) = \text{Binom}(e^n, \theta).$$

It would be tempting to consider $w_{ij}\mathbf{1}\{z_i = \mathbf{z}_j\}$ as individual Bernoulli variables; however, this would require conditioning on the indices for which $w_{ij} = 1$, leaving no other source of randomness (as we already condition on the partition $\rho$). Instead, consider the $n$th areal unit, with $d_n$ connections to other units ($d_n$ is also the *degree* of unit $n$), indexed by $i_1, i_2, \ldots, i_{d_n}$. Each of these indices can be assigned to a unit in the same cluster as $n$, and thus represent a *within-cluster* connection, or to one in a different cluster. Each index $i_k$ can thus identify the Bernoulli random variable $Y_k^n$ which is used to define $w$:

let $Y_k^n = \mathbf{1}\{z_{i_k} = z_n\}$. If we assume that $Y_1^n, \ldots, Y_{d_n}^n$ are independent and identically distributed $Y_k^n \sim \text{Bern}(\theta)$, then $v_n = \sum_{k=1}^{d_n} Y_k^n$ is the number of within-cluster connections of unit $n$. Moreover, let $v^n = \sum_{i \leq n} v_i$ be the number of within-cluster connections generated by the first $n$ areal units, then $v^n = v^{n-1} + v_n$. To complete the auxiliary model specification we assume $\theta \sim$Beta$(a, c)$; see Appendix C.1 for a heuristic to specify the hyper-parameters. Then

$$g(\mathbf{x}^n) = q(v^n) = \binom{v^n}{e^n} \frac{B(a + v^n, c + e^n - v^n)}{B(a, c)}. \tag{4.7}$$

Note: the assumption of independence is only approximate, because the indices are not sampled with replacement; however for simplicity we consider this approximate distribution.

One of the important properties of the PPMx is the coherence across sample size; Müller et al. (2011) show that this property holds when we condition on the full data $\mathbf{x}^n$. When instead of conditioning on the full data we condition on a summary statistics, the property does not hold exactly in the same way:

$$\begin{aligned} p(\rho_n | v^n) &= \sum_{z_{n+1}} \int p(\rho_n, z_{n+1}, v_{n+1} | v^n) dv_{n+1} = \\ &= \sum_{z_{n+1}} \int p(\rho_{n+1} | v^n, v_{n+1}) p(v_{n+1} | v^n) dv_{n+1} \end{aligned}$$

The difference that we notice is that in this formula we have $p(\rho_{n+1} | v^n, v_{n+1})$ instead of $p(\rho_{n+1} | v^{n+1})$, which would be more desirable.

A possible way to fix this would be to consider the sequence of $(v_i)$ for $i = 1, \ldots, n$: $p(\rho_n | v_1, v_2, \ldots, v_n)$, where $\sum_{i=1}^n v_i = v^n$. This would imply conditioning on the sequence of summary statistics, which represents more information and does not have any apparent practical benefits or intuition; moreover, conditioning on the sequence would make the cohesion depend on the order the areal units are observed, which is not desirable.

## 4.4. Comparison of prior properties

In this section we outline and compare some properties of the distributions for partitions of areal data, described in the previous sections: (1) the Dirichlet Process, (2) the Dirichlet Process restricted on connected partitions (connected-DP), (3) the EPA (Dahl et al., 2017), (4) the DP-MRF (Orbanz and Buhmann, 2008), (5) the SBM-PPMx and (6) the Areal-PPMx, both described in Section 4.3. We consider their *prior* properties, i.e. the properties induced on the partitions when these distributions are used as prior distribution, without data being observed. Knowing the prior properties is important, as the choice of the prior distribution often influences the posterior inferences. Moreover practitioners should be aware of these properties, as they might want to choose a distribution that describes their prior beliefs.

The properties we consider can be divided in two categories: the ones relevant for any distribution over partitions and the ones specifically relevant for areal data. In the first category we consider (a.) the distribution of the number of clusters, (b.) the distribution of the size of the largest cluster and (c.) the distribution of the number of singletons. In the second category we consider (d.) the distribution of the number of within-cluster connections and (e.) the proportion of connected partitions.

For simplicity we consider areal data arranged in regular square grids. Since the number of partitions grows more than exponentially, it is not feasible to exactly calculate these properties for grids larger than the 3 by 3 grid. Thus we report the exact properties for such graph, and we estimate them for larger graphs, by sampling partitions from these prior distributions. While it is easy to sample from the Dirichlet Process and the EPA priors, it is impossible to directly sample from the other prior distributions that involve a cohesion function (the connected-DP, the DP-MRF, the PPMx-SBM and the Areal-PPMx). For such distributions we will use importance sampling: we draw partitions from the DP or EPA distribution and re-weight the samples accordingly.

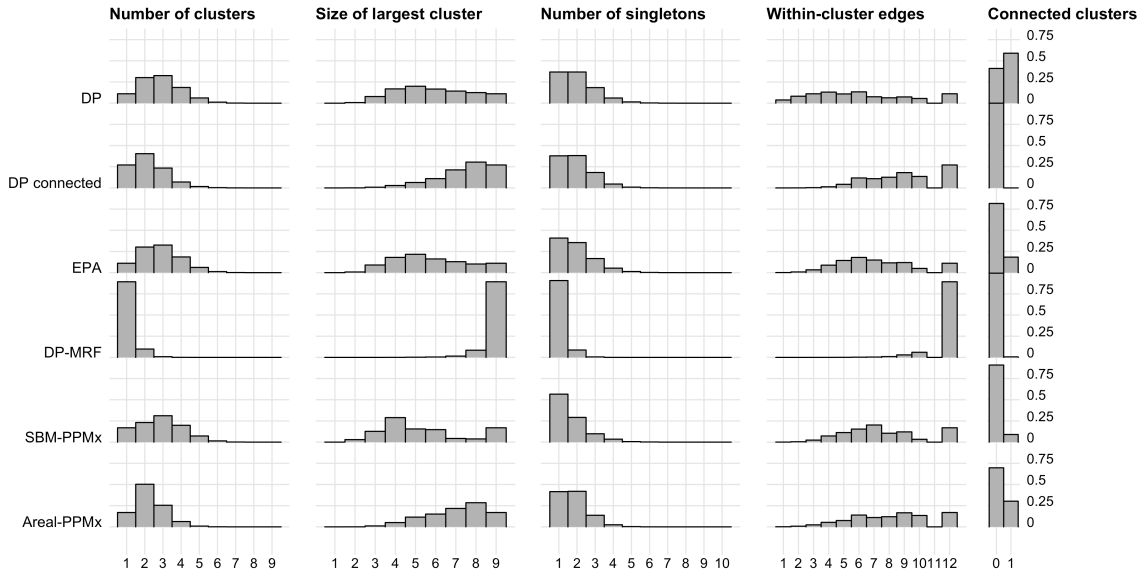


Figure 15: Distribution of the properties (a)-(e) on a 3 by 3 grid.

In figures 15 we report the properties (a)-(e) computed exactly on the 3 by 3 grid graph. We notice how the EPA distribution has behaviors extremely similar to the DP, with the exception of (d) the number of within-cluster edges and (e) the proportion of connected partitions, where we see that it favors the spatial clusters compared to the DP; however, compared to the other distributions, the EPA is the most similar to the DP in the number of within-cluster connections (d), even though it seems to induce a moderately large proportion of connected partitions (e). On the contrary, we notice that the DP-MRF is the distribution that mostly differs from the DP: the MRF cohesion function puts so much weight on the partition with one cluster that all the properties are completely distorted in favor of such partition. The Connected-DP prior displays properties that are similar to the DP but with a reasonable preference for spatial partitions. However, this distribution does not allow for partitions with a disconnected cluster, which is quite a restrictive assumption. The Areal-PPMx seems to display properties similar to the Connected-DP, with the exception of the proportion of connected partitions (e), where it allows a positive proportion of non-connected partitions, even though much smaller than the DP. The SBM-PPMx displays properties apparently similar to the DP, while giving higher weight to connected partitions

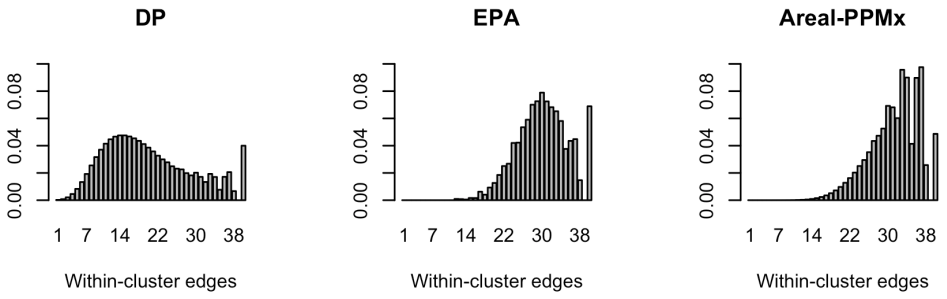


Figure 16: Distribution of the number of within-clusters connections for different distributions, computed approximately using importance sampling on a 5 by 5 grid.

or partitions with a larger number of within-cluster connections.

In the section C.2 of the appendix, we report the same properties computed approximately on a 5 by 5 grid graph, using importance sampling. Here, we simply emphasize some important points.

First, there is a strong difference in the distribution of the number of within-cluster connections between the EPA and the Areal-PPMx, and it is even stronger than in the 3 by 3 grid. In figure 16 we report a zoomed-in plot, where it is possible to see more clearly that compared to the DP, the EPA gives some more weight to partitions with a large number of within-cluster connections; however such weight is not very strong compared to other distributions, as the Areal-PPMx.

One might wonder whether it's possible to increase the strength of the pairwise similarity function to induce a stronger spatial effect. To that purpose, we examine the distribution of within-cluster connections for the exponential similarity function $\lambda(i,j) = \exp(-\tau \cdot d(i,j))$, for 5 different values of $\tau$: $\tau = 0$, which corresponds to the DP, $\tau = 1, 3, 10$ and 50. In figure 17 we can see that while the "spatial effect" increases from the DP to larger values of $\lambda$, such as 3 and 10, we also notice that for very large values of $\lambda$, such as 50, the effect stops increasing. As for the Areal-PPMx, it is instead possible to tune the values of the hyper-parameters to make the spatial effect much stronger.

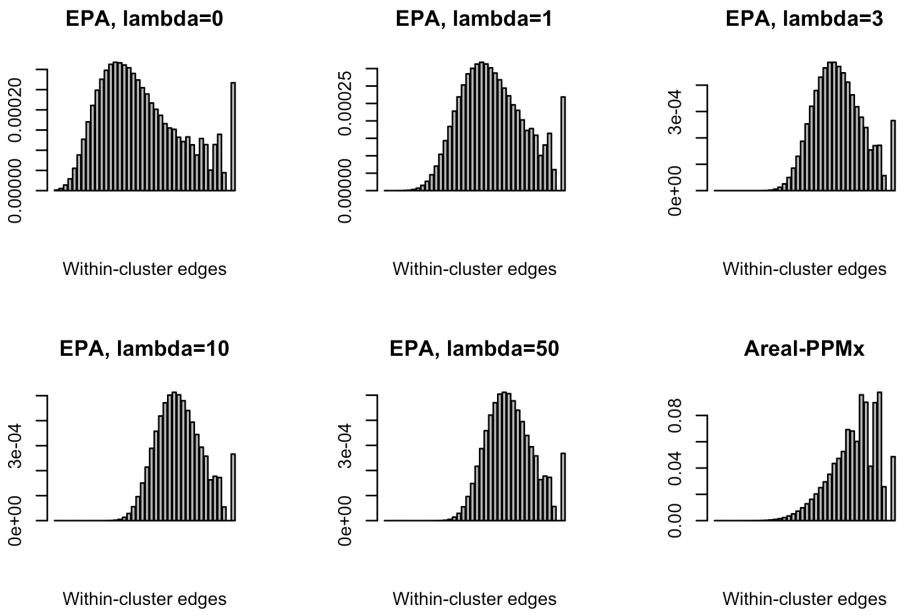


Figure 17: Distribution of the number of within-clusters connections for the DP, the EPA with different $\tau$ values and the Areal-PPMx, computed approximately using importance sampling on a 5 by 5 grid.

A second notable point is that the SBM-PPMx displays properties that differ from the DP in unexpected ways. In particular, the distribution of the number of within-cluster connections (d) is quite spiky compared to EPA and Areal-PPMx distributions, which are somewhat flatter. Instead, if we compare it to the Connected-DP or the DP-MRF, which also have an important mode, we notice that the most frequent number of within-cluster connections is much smaller than the total number of connections, which is the mode for the Connected-DP or the DP-MRF. We believe this is due to the difficulty in tuning the hyper-parameters of the SBM-PPMx (a heuristic for specifying them is described in section C.1 of the appendix). In particular, the effect is stronger for larger graphs, such as the 5 by 5 grid. In particular, it is not trivial to find a balance between incorporating sufficient prior information to induce the desired spatial behavior and distorting the distribution by giving too much weight to a specific type of configuration. In the simpler auxiliary model of the Areal-PPMx prior this problem fortunately did not appear and the behavior seems to be consistent with the analysis on the smaller graph.

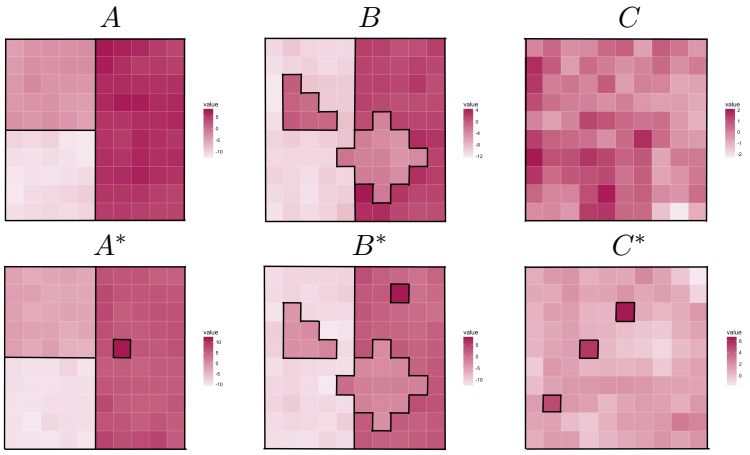


Figure 18: Partition configurations considered in the analysis of posterior properties. Configurations denoted with a star are similar to the corresponding configuration without a star, with the addition of some singletons.

## 4.5. Comparison of posterior properties

When the prior distribution is combined with the data model, and the posterior distribution is considered, it is possible to make inferences. In this section we analyze the posterior properties and behaviors for the distributions described in the previous sections, except for the SBM-PPMx distribution.

For this purpose we consider different partition configurations, reported in figure 18, on a 10 by 10 grid graph. We analyze three types of configurations with different numbers and shape of clusters: configurations $A$, $B$ and $C$; for each type we also consider a version that also displays some singletons: configurations $A^*$, $B^*$ and $C^*$. Considering partitions with singletons is important because they have fewer within-cluster connections, as the singleton cluster is formed by one unique element, while still agreeing with the concept of being "spatial". Moreover, as many distributions for spatial partitions prefer large numbers of within-cluster connections, partitions with singletons risk to be penalized.

Each of these configurations was used to generate data with the following model: cluster-specific means are fixed chosen so that the different clusters are distinguishable; within each

cluster, unit-specific values are sampled using a conditionally auto-regressive (CAR) model (using the formulation from Leroux et al. (2000)) centered around the cluster-specific mean. In this simulated analysis, 300 datasets are sampled for each partition configuration.

We analyze the posterior distribution induced by each of the priors considered. We use the Particle Optimization algorithm to find the three partitions with largest posterior probability; we then consider the proportion of times the partition used to generate the data is found among the top three partitions.

| | Configurations | | | | | |
|---|---|---|---|---|---|---|
| Priors | $A$ | $A^*$ | $B$ | $B^*$ | $C$ | $C^*$ |
| DP | 0.82 | 0.71 | 0.47 | 0.37 | 0.99 | 0.75 |
| Connected-DP | **1** | 0.64 | **0.66** | 0.28 | **1** | 0.31 |
| EPA | 0.87 | 0.52 | 0.09 | 0.09 | 0.92 | 0.61 |
| DP-MRF | **1** | 0.83 | **0.66** | 0.34 | **1** | 0.55 |
| ArealPPMx | 0.91 | **0.86** | 0.52 | **0.48** | **1** | **0.75** |

Table 3: Proportion of times the original partition was recovered among the top three partitions using the Particle Optimization algorithm.

Table 3 reports the results for the configurations and for the different distributions. We note how the DP is not bad at recovering the original partitions, and even though it does not give higher probability to spatial partitions it can recover them if the data displays that behavior. The Connected-DP is better than the DP in all the configurations without singletons, but performs worse in configurations with them. The EPA performs particularly poorly in many of these configurations. In particular we found that it would recover finer partitions, picking up subtle data behaviors, similar to what a uniform distribution over partitions would find. For this reason, we believe the EPA can be considered as a non-informative or vague distribution, as in the posterior, the data likelihood would have much more weight than the prior. The DP-MRF seems to perform well in some settings and worse in others. The Areal-PPMx seemed to be particularly sensitive in recovering singletons, while also performing well in the other configurations.

## 4.6. Discussion

Different priors for distributions of areal data have been proposed, but the properties induced are known only for few of them. Under the Ewens-Pitman Attraction, the asymptotic number of clusters grows as for the Ewens-Pitman distribution. The Product Partition Model dependent on covariates (PPMx) is coherent across sample sizes; however, though it has been adapted to spatial point-referenced data, it has not been extended to the case of areal data. The properties of other distributions, like the DP restricted on connected partitions and the Markov Random Field DP, have not been explored.

In this work, we compared the prior and posterior properties of several distributions that can be used to model partitions of areal data. Moreover, we introduced two auxiliary models that extend the PPMx to partitions of areal data. The first, the SBM-PPMx, models the adjacency structure as a network and is based on the Stochastic Block Model (SBM). The second, the Areal-PPMx, models a summary statistic of the adjacency information, the number of within cluster connections, using a Beta-Binomial distribution.

Most of the distributions we consider have been constructed by modifying the Dirichlet Process distribution to induce spatial behaviors. However, in our analyses of prior properties we showed that for some of them, such as the MRF-DP, the properties displayed are strongly distorted from the properties of the DP. Others, such as the EPA, display very similar properties to the DP, but are not able to incorporate arbitrarily high level of spatial dependence. The connected-DP displays good spatial properties, but imposes a hard constraint and gives zero probability to partitions with non connected clusters. The properties for the SBM-PPMx strongly depend on the hyper-parameter specification, which can be complicated, especially for larger number of units. The Areal-PPMx instead displays good properties, similar to the EPA, with adaptable level of spatial dependence; the hyper-parameter specification for this model seemed to not influence the prior as much. However, the Areal-PPMx does not have the property of coherence across sample sizes, despite being

constructed with an auxiliary model, because it depends on a summary statistic of the data, and not individual data points.

In a similar way, these same distributions can be adapted to modify the Ewens-Pitman distribution, for which the number of clusters is characterized by a power-law behavior, rather than the logarithmic growth of the Ewens distribution. In future work, additional analyses can be performed, to study their prior properties under this new setting and to test whether the power-law behavior is preserved.

One of the distributions proposed in this work, the SBM-PPMx, suffers from bad empirical properties and sensitivity to hyper-parameters choice. Several reasons can possibly explain this behavior. First, the likelihood chosen for the auxiliary model is designed for graphs, and while areal data can be interpreted as a graph, they represent a particular kind of graphs: planar graphs, i.e. graphs in which edges do not intersect. It is likely that a model explicitly constructed for planar graphs would have a better performance than the Stochastic Block Model. And second, the SBM is a complex model that depends on several hyper-parameters, and its success depends on the correct specification of these values, which are usually estimated, not specified a priori. Since the auxiliary model is used to specify a prior distribution, it needs to be a simple model, with a small number of hyper-parameters that can be simply be specified without negatively affecting the behavior of the prior distribution. The other distribution we proposed, the Areal-PPMx, achieves this goal by modeling a summary statistic of the data, which requires a simpler auxiliary model, with easier tuning of hyper-parameters. However, conditioning on a summary statistic changes the structure of the PPMx framework, and the property of coherence across sample sizes is lost. Future work should focus on constructing a different auxiliary model that combines the positive aspects of the SBM-PPMx and Areal-PPMx. To ensure the property of coherence across sample sizes guaranteed by the PPMx framework, it should model unit-specific information, instead of a summary statistic of all the units; to provide good empirical behavior, it would need to be simple and not sensitive to hyper-parameter choice.

# Chapter 5

# Clustering Data at Multiple Resolutions

## 5.1. Introduction

Clustering is a popular unsupervised method that has been studied widely in statistics and machine learning. In particular, when data are divided into a large number of categories, it can be beneficial to cluster together those displaying similar observations. In fact, clustering not only improves interpretability of the categories that are grouped together, but it also reduces the dimensionality of the problem. Categories are often organized in a hierarchy, with coarser and finer classes. Examples range from image recognition, such as an image being classified as a "fruit" or an"apple," to diagnosis codes, such as patients being admitted to a hospital for a "lung problem" or an "infection by pencillin-resistant *Strep. pneumoniae*," to topic modeling, such as an article talking about "sport" or "football." This hierarchical structure has generated a great deal of interest in different disciplines, including in settings different from clustering (see for example Yan et al. (2015); Singh et al. (2014); Blei et al. (2010)).

A notable example of hierarchically structured categories is represented by geo-referenced data that are aggregated within regions. These aggregated observations are called *areal data* and the corresponding regions are *areal units*. Often multiple levels of granularity exist, from fine resolutions such as city neighborhoods, to coarser ones such as counties or states. Because each fine unit is contained in a coarse one, a hierarchy is induced on the regions and as a consequence on the areal data collected at different resolutions. In this way, each level of the hierarchy corresponds to a granularity layer.

In previous research, many studies have focused on clustering areal data, especially in the domain of disease mapping (see e.g. Knorr-Held and Raßer, 2000; Feng et al., 2016; Denison

and Holmes, 2001; Anderson et al., 2017), but also in the context of crime modeling (Balocchi et al., 2019). Clustering allows the creation of more interpretable maps, with areal units being grouped together when they display similar behaviors in the phenomena of interest. It also aids with dimensionality reduction to prevent over-parametrization and with flexible sharing of information between regions to improve parameter estimation. The majority of such analyses fix one layer of resolution and only perform the analysis at the chosen granularity. Often the highest possible resolution is chosen to better capture local patterns, even though it is more computationally intensive than lower resolutions.

However, in complex environments, such as cities, the spatial variation of the data can arise at different resolutions in different regions. It is known in fact that crime rates are higher in central than suburban areas (Zenou, 2003) and that the majority of the spatial variability happens at a micro level (Schnell et al., 2017; Steenbeek and Weisburd, 2016). It is unclear whether the spatial variability can be considered constant throughout the city, or whether there exists differences between central and suburban areas. For example, it is possible for the spatial distribution of crime in a city to be more uniform in a residential neighborhood compared to downtown areas, in which higher heterogeneity of the urban environment could be expected to cause more variation in crime behaviors. Standard clustering methods that consider a single granularity level might not be able to capture those differences. Therefore, instead of reducing the analysis to a specific level, it can be beneficial to simultaneously consider multiple resolutions.

In this work we focus on clustering crime behaviors at multiple resolutions in the City of Philadelphia. By finding clusters of neighborhoods with similar crime levels, we can flexibly model crime behaviors and simultaneously reduce the dimensionality and produce interpretable maps. Moreover, by modeling crime aggregated at multiple resolutions, we can assess whether one level is suitable to explain the variation of crime, or whether distinct levels are more appropriate to describe that variation in different parts of the city. We consider two different subdivisions of the city. We first aggregate crime at the Police Division

and Police Service Area resolutions. We then consider the census tract and block group resolutions for a subset of the city corresponding to the area of West Philadelphia.

When the number of clusters is unknown *a priori*, Bayesian nonparametric models are convenient choices as they can specify a prior distribution over the space of partitions for which the number of clusters does not need to be fixed. While the Dirichlet Process is a popular model for clustering in the Bayesian nonparametric framework, it can only be used to cluster data within one level of resolution. For multi-resolution clustering, the nested Dirichlet Process (Rodríguez et al., 2008) is applicable because it considers grouped data and partitions both the observations and the groups themselves. Thus, if used to cluster crime in Philadelphia, it could simultaneously find partitions of the census tracts and of the PSAs, if the former are regarded as observation units and the latter as groups.

In the partitions generated by the nested Dirichlet process however, clusters at finer levels are forced to be nested within the clusters at coarser levels, as shown in Camerlenghi et al. (2018). This characteristic is quite restrictive in our modeling, as distinct behaviors can be found at the intersection of coarser level units. To allow for more flexible partitions which are not forced to be nested, in this work we introduce a model, the nested Hierarchical Dirichlet Process (nHDP), that extends the nested Dirichlet process by incorporating it with the Hierarchical Dirichlet Process (Teh et al., 2006).

Some similarities to this line of work are present in the work of Blei et al. (2010) and Paisley et al. (2014) on *hierarchical topic models*. In particular, the *nested Chinese Restaurant Process* (nCRP) (Blei et al., 2010) generates a distribution over infinite trees of topics, which is combined with the modeling of words in text documents for detecting a hierarchy of topics. The nCRP could be viewed as an infinite-level extension of the nested Dirichlet Process, but its use is substantially different than the mixture models proposed by Rodríguez et al. (2008). The *nested Hierarchical Dirichlet Process* of Paisley et al. (2014) extends the nCRP by creating a hierarchy of infinite trees of topics, to allow for document-specific topic trees. Thus this model, while it has the same name as the model proposed by our work, is essen-

tially different and is not suitable for multi-level clustering. More similarly related to our model is the work developed by Agrawal et al. (2013); Tekumalla et al. (2015); they present a *nested HDP* model, which differs from our proposal but includes our model as a special case. However, their application is different from the one considered in this work: their focus is on entity-topic models (that can be seen as a specific instance of multi-level modeling in the context of text data); moreover, their proposed sampling algorithm differs from ours. Additionally, a similar idea was independently developed by Giovanni Rebaudo and collaborators (personal communication), although their work is focusing on more theoretical aspects, it considers a different application and a different sampling approach; moreover, a similar but different model has been simultaneously developed by Denti (2020).

The rest of this chapter is organized as follows: before presenting our model in section 5.3, we review some Bayesian nonparametric models in section 5.2. In section 5.4 we describe the crime data and the models for the multi-resolutions analysis; we study the performance of our mode with simulated data and apply it to the Philadelphia crime data.

## 5.2. Background

The Dirichlet Process (Ferguson, 1973) is a distribution over random probability distributions; it is characterized by a concentration parameter $\alpha > 0$ and a base distribution $H$. A realization $G$ from the Dirichlet Process is almost surely discrete and can be written as $G = \sum p_k \delta_{\theta_k^*}$, where $\delta_{\theta_k^*}$ are the atoms of $G$ and $p_k$ is the probability associated with $\theta_k^*$. According to the *stick-breaking construction* of $G$ (Sethuraman, 1994), the atom locations $\theta_k^*$ are i.i.d. random variables distributed according to $H$ and the probabilities $p_k = b_k \prod_{j=1}^{k-1}(1 - b_j)$, where $b_j \overset{iid}{\sim} \text{Beta}(1, \alpha)$; we will denote the stick-breaking construction with $(p_k) \sim SB(\alpha)$.

Because of its discreteness, the Dirichlet Process is rarely employed to directly model data; instead, it is mostly used for specifying a prior distribution for the mixture components in mixture modeling. Let $\{x_1, \ldots, x_n\}$ be a set of exchangeable observations drawn from a

mixture model, and let $\{\theta_1, \ldots, \theta_n\}$ be the latent mixture components associated with each observation: $x_i$ given $\theta_i$ is drawn from $F(\theta_i)$. To specify a prior distribution on the latent mixture components we assume $\theta_1, \ldots, \theta_n | G \overset{iid}{\sim} G$ and $G \sim DP(\alpha, H)$. We then say that $x_1, \ldots, x_n$ follow a *Dirichlet Process mixture* model. For the rest of this description we will mostly focus on the modeling of the latent mixture components, and we will assume that the observations are drawn from a parametric distribution $F$ parametrized by the mixture component.

The discreteness of $G$ also implies that among the latent mixture components $\theta_1, \ldots, \theta_n$ some values will be repeated with high probability. Therefore a partition $\gamma$ is induced on the data by the mixture components, with clusters identified by their unique values $\theta^*_{i_1}, \ldots, \theta^*_{i_K}$: for $k = 1, \ldots, K$, we define $S_k = \{i : \theta_i = \theta^*_{i_k}\}$ and $\gamma = \{S_1, \ldots, S_K\}$. When $G$ is drawn from a Dirichlet Process, the partition $\gamma$ is exchangeable and is distributed according to the *Chinese Restaurant Process* (Aldous, 1985).

However, often the observations cannot be assumed to be exchangeable because they are divided into known groups: consider $m$ groups of data, $\{x_{j1}, \ldots, x_{jn_j}\}$ for $j = 1, \ldots, m$, and the corresponding latent mixture components $\{\theta_{j1}, \ldots, \theta_{jn_j}\}$. The Hierarchical Dirichlet Process (HDP) (Teh et al., 2006) is an extension of the Dirichlet Process that can model grouped data, so that clusters can be shared across groups. This is achieved by considering group-specific discrete measures $G_j$ that are realizations from a Dirichlet Process with a discrete base measure $G_0$, where $G_0$ is itself a realization from a Dirichlet Process. The mixture components for group $j$ are then sampled from $G_j$:

$$
\begin{aligned}
G_0 &\sim DP(\alpha, H) \\
(G_j)_{j=1}^m | G_0 &\overset{iid}{\sim} DP(\alpha_0, G_0) \\
(\theta_{jl})_{l=1}^{n_j} | G_j &\overset{iid}{\sim} G_j.
\end{aligned}
$$

Since the $G_j$ are realizations from a Dirichlet Process with discrete measure $G_0$, all the $G_j$

share the atoms of $G_0$: $G_j = \sum_k p_{jk}\delta_{\theta_k^*}$, for $j = 0, 1, \ldots m$ (see Teh et al. (2006) for the stick breaking construction for $(p_{jk})$). Thanks to the shared atoms of the $G_j$, observations in different groups can be assigned to the same cluster; in fact, there is a positive probability that data in two different groups $j_1$ and $j_2$ are assigned to the same component $\theta_k^*$ (the probability is given by $p_{j_1k}p_{j_2k}$), for all $k$, since $p_{jk} > 0$ for all $j, k$. Figure 19 shows a graphical representation of the HDP model. Under the HDP, the distribution on the partition of all the data aggregated across groups is called the *Chinese Restaurant Franchise*. While this distribution is appropriate in the presence of grouped data, it does not capture any similarities that might be present in various groups.

When the data is divided in groups, but we also expect some of the groups to display the same behavior, the nested Dirichlet Process (nDP) (Rodríguez et al., 2008) can be useful. This model can be described as a Dirichlet Process in which the base distribution is another Dirichlet Process. It is characterized by the parameters for the base process, the concentration parameter $\alpha > 0$ and the base distribution $H$, and an additional concentration parameter $\beta > 0$. We can informally write that a realization $Q$ from the nested Dirichlet process ($Q \sim nDP(\beta, \alpha, H)$) is sampled from $Q \sim DP(\beta, DP(\alpha, H))$ and, more precisely, $Q = \sum w_i\delta_{G_i^*}$, where $(w_k) \sim SB(\beta)$, and $G_i^* \overset{iid}{\sim} DP(\alpha, H)$. Since the atoms of $Q$ are independent realization from a Dirichlet Process, we can write them as $G_i^* = \sum_k p_{ik}\delta_{\theta_{ik}^*}$, where $(p_{ik})_k \sim SB(\alpha)$ and $\theta_{ik}^* \overset{iid}{\sim} H$:

$$
\begin{aligned}
Q &= \sum w_i\delta_{G_i^*} \\
(w_k) &\sim SB(\beta) \\
G_i^* &\overset{iid}{\sim} DP(\alpha, H).
\end{aligned}
$$

One of the unique properties of the nested Dirichlet Process is that it allows simultaneous clustering of not only the observations but also of the groups themselves. For each group $j$, consider the variable $G_j$, sampled from $Q$, and conditionally on $G_j$, the mixture components

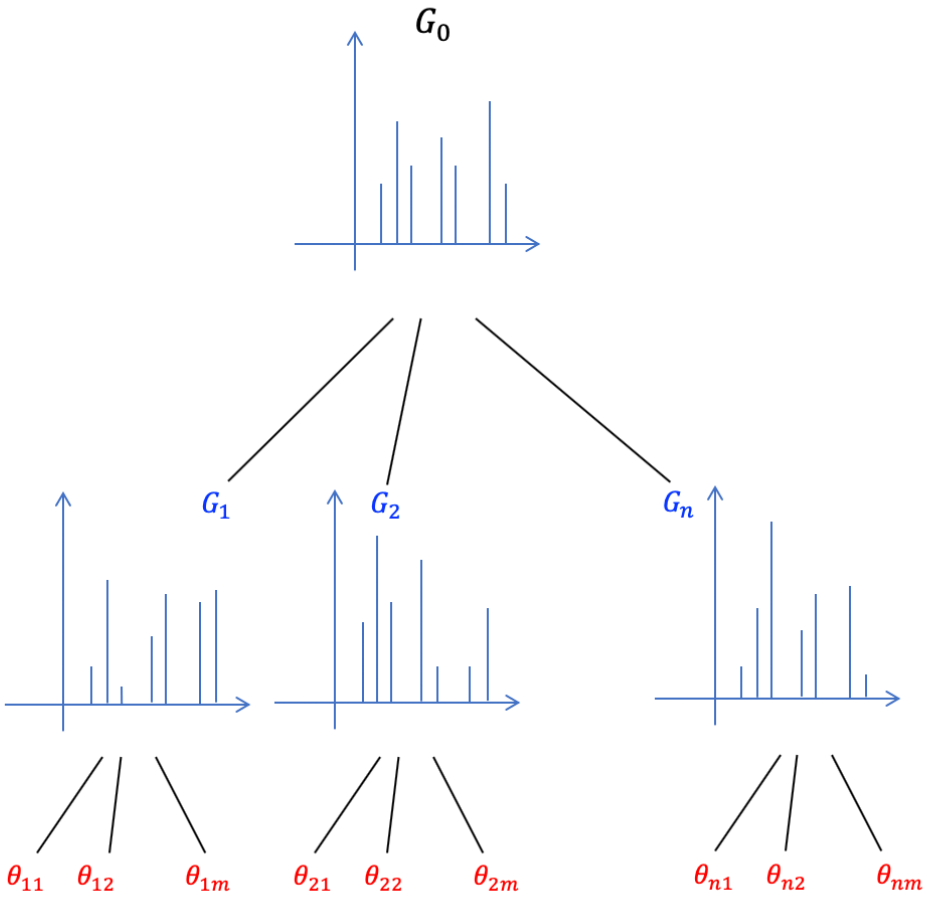


Figure 19: Diagram of the HDP model. Starting from the top, we represent a possible realization of $G_0$, which affects the distribution of the group specific $G_j$, which affect the distribution of the parameters $\theta_{jl}$. The graphical depiction of each discrete distribution uses vertical lines to represent the atoms of the distribution: the location of each line represents the location of the atom $\theta_k^*$, and the height of each line represents the atom's weight or probability. For the purpose of the plot, only a finite number of atoms are depicted. Note that since each $G_j$ shares the same atoms of $G_0$, the location of the vertical lines is the same for all the distributions, but the height of the lines varies. Since all the $G_j$ share the same atoms, the $\theta_{jl}$ across different groups can take on the same values and thus clusters can be shared across groups.

$\theta_{jl}$ are sampled from $G_j$. More formally,

$$
\begin{aligned}
Q &\sim nDP(\beta, \alpha, H) \\
(G_j)_{j=1}^{m} | Q &\overset{iid}{\sim} Q \\
(\theta_{jl})_{l=1}^{n_j} | G_j &\overset{iid}{\sim} G_j.
\end{aligned}
$$

In other words, the group specific variables $G_j$ identify the prior distribution of the group specific mixture components $(\theta_{jl})_{l=1}^{n_j}$ and since $Q$ is discrete, different groups can have the same distribution with positive probability. If $G_{j_1} = G_{j_2} = G_h^*$ for some $h$, then the observations in the two groups are identically distributed and the groups $j_1$ and $j_2$ are clustered together. Moreover, since the $G_i^*$ are discrete, the $\theta_{jl}$ are also divided into clusters, identified by the unique values $\theta_{ik}^*$. Figure 20 reports a graphical representation of the nDP and the model just described.

With the nested Dirichlet Process, two partitions can be considered: the partition of the groups, identified by the unique values of the $G_j$ and denoted with $\gamma^{(L)}$, and the partition of the observations, identified by the unique values of the $\theta_{jl}$ and denoted with $\gamma^{(H)}$.

As shown by Camerlenghi et al. (2018), the two partitions induced by the nested Dirichlet Process are nested: a cluster in the partition $\gamma^{(H)}$ contains observations from different groups only if such groups are clustered together in $\gamma^{(L)}$. The reason for this behavior lies in the fact that the atoms of $Q$ are independent realizations from a Dirichlet Process; in fact when two groups are not in the same cluster, i.e. $G_{j_1} = G_i^* \neq G_{j_2} = G_{i'}^*$, the atoms of $G_i^*$ and of $G_{i'}^*$ are going to almost surely differ from each other since $\theta_{ik}^* \overset{iid}{\sim} H$ and $H$ is a non-atomic distribution (thus there is zero probability that $\theta_{ik}^*$ is equal to $\theta_{i'k'}^*$). It follows that if the atoms of $G_{j_1}$ and $G_{j_2}$ are different with probability one, then the mixture components of the two groups cannot share values: $\theta_{j_1 l} \neq \theta_{j_2 l'}$ a.s.

This property is quite restrictive and non-desirable in multi-resolution clustering, and specifically in our application to clustering neighborhoods. In fact, two adjacent PSA could belong

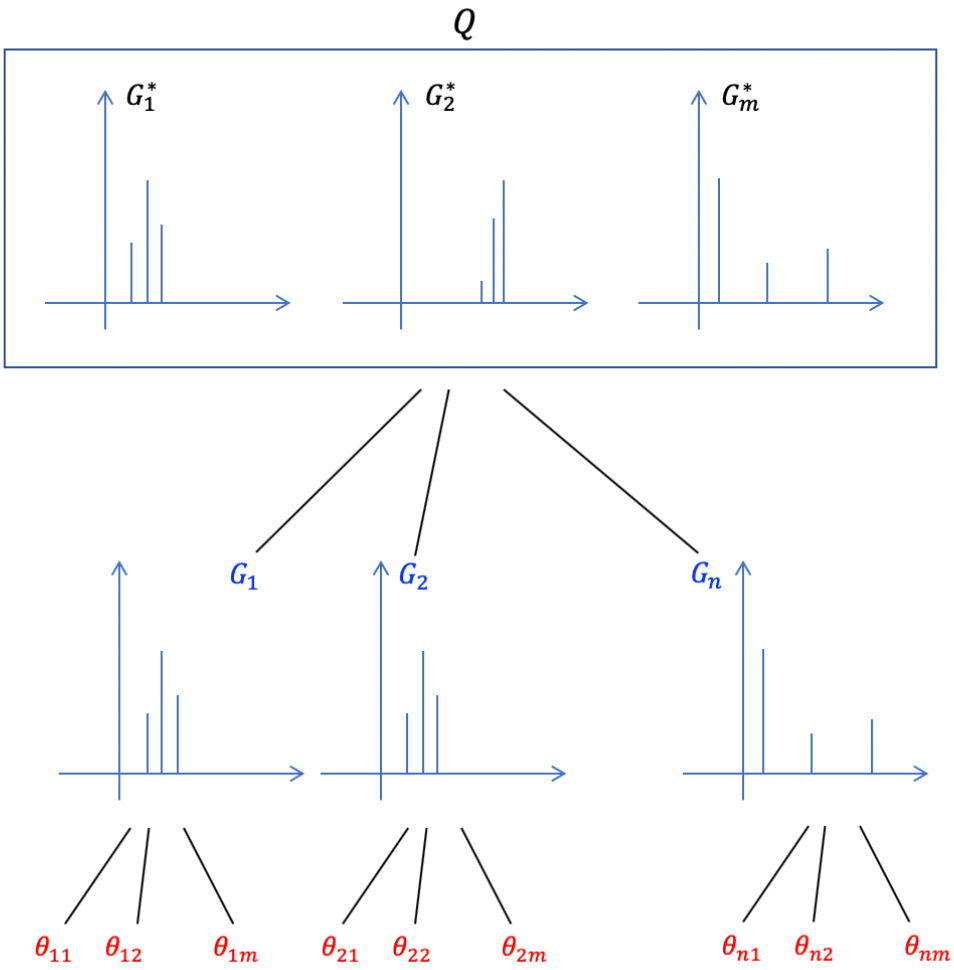


Figure 20: Diagram of the nDP model. Starting from the top, we represent a possible realization of $Q$, which affects the distribution of the group specific $G_j$, which affect the distribution of the parameters $\theta_{jl}$. The discrete distribution $Q$ is represented as a rectangular box containing other discrete distributions as its atoms, the $G_i^*$. The graphical depiction of each discrete distribution uses vertical lines to represent the atoms of the distribution: the location of each line represents the location of the atom, and the height of each line represents the atom's weight or probability. For the purpose of the plot, only a finite number of atoms are depicted. Note that different $G_i^*$ have different atoms (the vertical lines are located in different points), and that the $G_j$ coincide with one of the atoms of $Q$, one of the $G_i^*$. The $G_j$ that are equal to the same $G_i^*$, such as $G_1$ and $G_2$ which are equal to $G_1^*$, will share the same atoms and the same weights (location and height of the vertical lines), thus the underlying $\theta_{jl}$ can take on the same values and share clusters across groups. However, if one $G_j$ is equal to a different $G_i^*$, such as $G_n$ which is equal to $G_m^*$, the atoms and weight are different and the underlying $\theta_{jl}$ will take on different values from the parameters in different groups.

to separate clusters because of overall different trends, but at the same time some census tracts at the boundaries could display a similar behavior and we would want to cluster them together and capture such finer resolution pattern. However this would not be possible using the nested Dirichlet Process, because under that model census tracts in different PSAs cannot be clustered together unless the corresponding PSAs are in the same cluster. To allow for more flexible partitions in the next section we combine the nested Dirichlet Process of Rodríguez et al. (2008) with the Hierarchical Dirichlet Process of Teh et al. (2006).

## 5.3. The Nested Hierarchical Dirichlet Process

To cluster observations that are divided into groups and different groups can display the same behaviors, the model should allow for both the observations and the groups to be clustered. Moreover, allowing clusters to be shared between different groups would increase the model flexibility. It is then natural to combine the nested Dirichlet Process with the Hierarchical Dirichlet Process; this approach has also been taken by Giovanni Rebaudo and collaborators (personal communication). Similarly to the nested Dirichlet Process, consider a discrete distribution $Q$ whose atoms are themselves discrete distributions: $Q = \sum w_i \delta_{G_i^*}$ and $(w_k) \sim SB(\beta)$; however, instead of sampling the atoms $G_i^*$ from another Dirichlet Process as in the nDP, they are realizations from a Hierarchical Dirichlet Process: $G_i^*|G_0 \sim DP(\alpha_0, G_0)$ and $G_0 \sim DP(\alpha, H)$. For simplicity we write $Q \sim nHDP(\alpha_0, \alpha, \beta, H)$:

$$
\begin{aligned}
Q &= \sum w_i \delta_{G_i^*} \\
(w_k) &\sim SB(\beta) \\
G_i^*|G_0 &\overset{iid}{\sim} DP(\alpha_0, G_0) \\
G_0 &\sim DP(\alpha, H).
\end{aligned}
$$

If we consider the group-specific variable $G_j$ and the mixture components $\{\theta_{j1}, \ldots, \theta_{jn_j}\}$, then each $G_j$ is sampled from $Q$ and conditionally on $G_j$ the mixture components $\theta_{jl}$ are

sampled from $G_j$:

$$
\begin{aligned}
Q &\sim nHDP(\alpha_0, \alpha, \beta, H) \\
(G_j)_{j=1}^m | Q &\overset{iid}{\sim} Q \\
(\theta_{jl})_{l=1}^{n_j} | G_j &\overset{iid}{\sim} G_j.
\end{aligned}
$$

As in the nested Dirichlet Process, two groups are clustered together when their group-specific variables coincide, i.e. $G_{j_1} = G_{j_2}$, and the corresponding observations are sampled from the same distribution. When the two groups do not belong in the same cluster, $G_{j_1} = G_i^* \neq G_{j_2} = G_{i'}^*$, then differently from the nDP, the corresponding observations can still be clustered together because $G_i^*$ and $G_{i'}^*$ share the same atoms and there is a positive probability that $\theta_{jl} = \theta_{j'l'}$.

This property is especially important for multiple resolution clustering, as it allows partitions to be flexibly recovered, without forcing them to be nested within each other. In addition, it is important to consider models such as the nHDP when there can be similarities between groups and some groups contain a small number of observations. In such cases, it is of great importance to share information between groups and thus benefit from the increased size that is granted by their clusters.

In Figure 21 we represent pictorially the nDP and HDP models and how they are combined to construct the nHDP: as the diagram shows, both the nDP and the nHDP share the discrete measure $Q$, but its atoms differ, with the atoms in the nHDP being equal to the measures in the HDP.

### 5.3.1. Partitions

Thanks to the discrete nature of the realizations from a Dirichlet Process, samples from these realizations can be divided into clusters, based on their unique values. In the nDP and nHDP two types of partitions are induced, because of the two levels of discrete distributions:

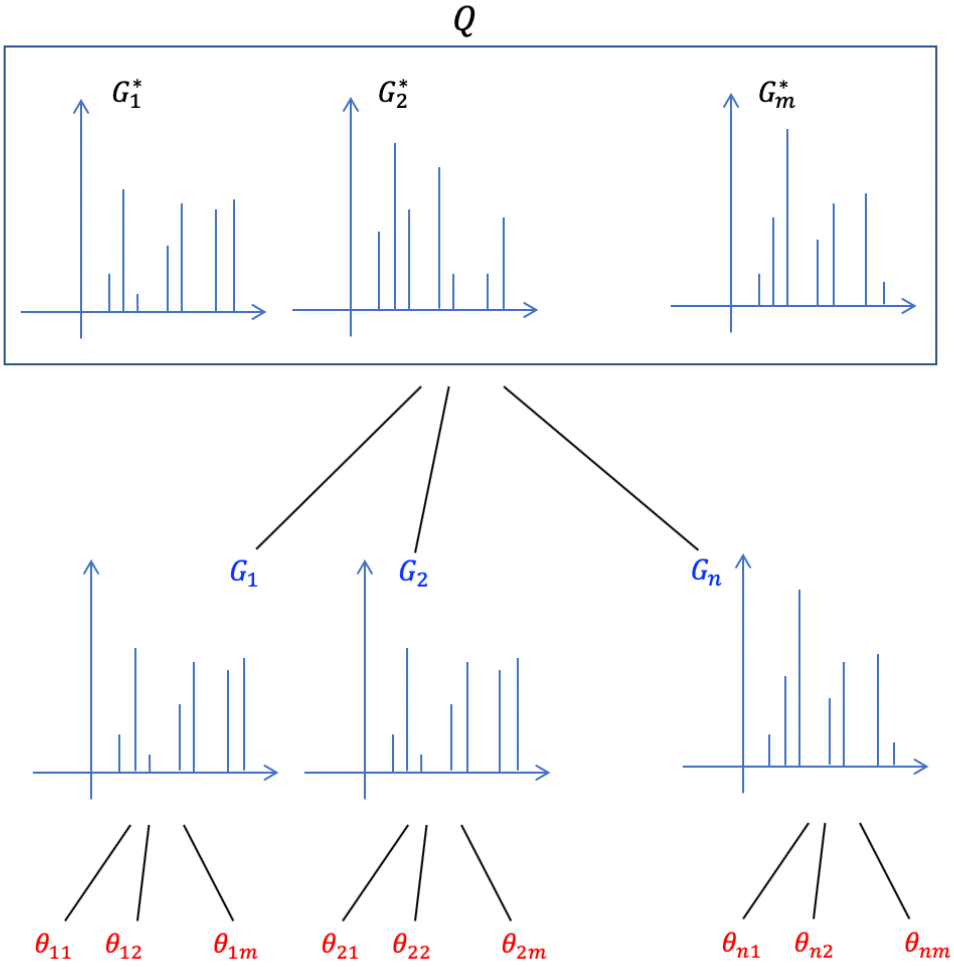


Figure 21: Diagram of the nHDP model. Starting from the top, we represent a possible realization of $Q$, which affects the distribution of the group specific $G_j$, which affect the distribution of the parameters $\theta_{jl}$. The discrete distribution $Q$ is represented as a rectangular box containing other discrete distributions as its atoms, the $G_i^*$. The graphical depiction of each discrete distribution uses vertical lines to represent the atoms of the distribution: the location of each line represents the location of the atom, and the height of each line represents the atom's weight or probability. For the purpose of the plot, only a finite number of atoms are depicted. Note that different $G_i^*$ have the same atoms but different weights (the vertical lines are located in the same locations but have different height), and that the $G_j$ coincide with one of the atoms of $Q$, one of the $G_i^*$. Even when two groups have different distributions, such as $G_1$ and $G_n$ which are respectively equal to $G_1^*$ and $G_m^*$, the atoms are the same, thus allowing the underlying $\theta_{jl}$ to share clusters even if their groups are not clustered together.

one on the groups and one on the observations. The partition on the groups, denoted with $\gamma^{(L)}$, is induced by $Q$ on the group-specific distributions $G_j$, i.e. the clusters $S_k^{(1)}$ are defined based on the unique values of the $G_j$: $S_k^{(1)} = \{j : G_j = G_{i_k}^*\}$. The partition on the observations $\gamma^{(H)}$ instead is induced by the collection of $G_j$ on the $\theta_{jl}$, and the clusters $S_h^{(2)}$ contain observation $l$ in group $j$ if its latent variable $\theta_{jl}$ is equal to the unique value $\theta_{k_h}^*$: $S_h^{(2)} = \{(j,l) : \theta_{jl} = \theta_{k_h}^*\}$.

We can think of these two partitions in a hierarchy: if we represent the relationship between groups and observations in a tree, where groups are the first level below the root and observations are in the next level, branching out from the groups nodes, like in Figure 21, then $\gamma^{(L)}$ is the partition of the first level of the tree and $\gamma^{(H)}$ is the partition of the second and lowest level.

We now consider the distributions induced on these partitions by the nHDP.

**Proposition 2.** *The marginal prior distribution induced by the nHDP$(\alpha_0, \alpha, \beta, H)$ on the partition of groups $\gamma^{(L)}$ is the Chinese Restaurant Process:*

$$p(\gamma^{(L)}) = CRP(\beta).$$

*Proof.* Conditional on $G_0$, $Q$ is a realization of a Dirichlet Process with base measure equal to a $DP(\alpha_0, G_0)$; this means that $Q$'s atoms $G_i^*$, despite sharing the same atoms $\theta_k^*$, are all different *a.s.*, because their sequence(s) of weights $(p_{ik})$ is(are) all different *a.s.*. Moreover $Q$'s weights $(w_k)$ are generated according to the stick-breaking construction with parameter $\beta$. Thus, conditional on $G_0$, the conditional distribution of $G_j$ given $G_1, \ldots, G_{j-1}$ follows the Pólya Urn scheme:

$$G_j | G_1, \ldots, G_{j-1}, G_0 \sim \frac{1}{j-1+\alpha} \sum_{l=1}^{j-1} \delta_{G_l} + \frac{\alpha}{j-1+\alpha} DP(\alpha_0 G_0).$$

From this follows that $p(\gamma^{(L)} | G_0) = CRP(\beta)$. To find the marginal distribution of $\gamma^{(L)}$ we need to integrate out $G_0$; for this purpose, note that $G_0$ only affects the distribution of a

new observation, i.e. $DP(\alpha_0 G_0)$. Then we just need to show that, marginally on $G_0$, a new observation $G^*$ is different from the previously observed $G_1, \ldots, G_{j-1}$. Note that for any $G_i$, $p(G^* = G_i|G_0) = 0$, because the Dirichlet Process is a non-atomic distribution on the space of probability measures. Since this is true for any $G_0$, $p(G^* = G_i|G_1, \ldots, G_{i-1}) = \int p(G^* = G_i|G_0)p(dG_0|G_1, \ldots, G_{i-1}) = 0$. □

**Proposition 3.** *The prior distribution induced by the $nHDP(\alpha_0, \alpha, \beta, H)$ on the partition of observations $\gamma^{(H)}$ conditional on the partition of groups $\gamma^{(L)}$ is a Chinese Restaurant Franchise distribution, where the groups are defined by the clusters of $\gamma^{(L)}$.*

*Proof.* Let $\gamma^{(L)}$ be the partition of groups, where each cluster gathers together the groups that take on the same value of the group specific distribution: for example let $S_k^{(1)} = \{j : G_j = G^*_{h_k}\}$ where $h_k$ corresponds to the index of the unique value of $Q$ associated with those $G_j$. Then the observations in the groups that belong to the same cluster share the same distribution: if we merge together in a vector the observations in the groups that belong to the same cluster, $\boldsymbol{\theta}_{S_h^{(1)}} = \{\theta_{j1}, \ldots, \theta_{jn_j} : j \in S_h^{(1)}\}$, then

$$\begin{aligned} \boldsymbol{\theta}_{S_h^{(1)}}|G^*_h &\overset{iid}{\sim} G^*_h, \forall h \\ G^*_{h_1}, \ldots, G^*_{h_K}|G_0 &\overset{iid}{\sim} DP(\alpha_0, G_0) \\ G_0 &\sim DP(\alpha, H). \end{aligned}$$

Thus, conditional on $\gamma^{(L)}$, we can divide the $\theta_{jl}$ into the collections defined by the clusters $S_h$ and they are distributed according to a Hierarchical Dirichlet Process. By considering the $\theta_{jl}$ that take on the same values across the collections we can define $\gamma^{(H)}$, whose distribution is described by the Chinese Restaurant Franchise with groups defined by the clusters of $\gamma^{(L)}$. □

### 5.3.2. Posterior sampling

In this section we describe the Markov Chain Monte Carlo sampling schemes for the Nested Hierarchical Dirichlet Process mixture model. For simplicity we consider a model where $F$ and $H$ are conjugate distributions, so that the mixture components parameters $\theta_k^*$ can be integrated out.

We consider a sampling scheme in which only the latent partitions at the group level $\gamma^{(L)}$ and at the observation level $\gamma^{(H)}$ are iteratively sampled. In a first step we sample $\gamma^{(H)}$ given the group level partition and the data, and as we will show this can be carried out with one of the posterior sampling schemes for the HDP, for example those described in Teh et al. (2006). The second step, to sample $\gamma^{(L)}$ given the observation level partition, requires a more complex procedure: since the dimensionality of the space of $\gamma^{(H)}$ changes when we change $\gamma^{(L)}$, this problem falls into the category of models described by Green (1995). However, the reversible-jump framework described by Green cannot be simply applied in this setting, because of the discreteness and complexity of the space of $\gamma^{(H)}$. Instead, we formulate this step by specifying a proposal that allows for movements in the product space of $(\gamma^{(L)}, \gamma^{(H)})$: we propose a move for $\gamma^{(L)}$ and conditional on this proposed value we propose a value for $\gamma^{(H)}$ that is consistent with the new configuration.

**Sampling $\gamma^{(H)}$.** Given the partition of the groups $\gamma^{(L)}$, we know how the observations are divided into groups in the hierarchy or, using the terminology from the Chinese Restaurant Franchise, we know how the costumers (the observations) are divided into restaurants (the clusters of the group partition). Specifically, for all the groups in cluster $S_k^{(1)}$, we can aggregate all of the observations together and they form a group in the new hierarchy or a

restaurant in the metaphor terminology.

$$
\begin{aligned}
(\theta_{jl} : j \in S_k^{(1)}) | G_k^* &\overset{iid}{\sim} G_k^* \\
G_k^* | G_0 &\sim DP(\alpha_0, G_0) \\
G_0 &\sim DP(\gamma, H).
\end{aligned}
$$

Thus, we can sample the partition of the $\theta_{jl}$ using the standard Chinese Restaurant Franchise sampling schemes: we use the Chinese Restaurant Franchise representation and sample the assignment of costumers into tables within each restaurant and the assignment of tables to dishes across restaurants (Teh et al., 2006). However, rather than exploring the space with local moves that sample each cluster assignment iteratively using Gibbs-type updates, we modify the partitions with broader 'split-merge' moves that improve mixing (Jain and Neal, 2004b). We extend the algorithm of Wang and Blei (2012) to perform split-merge moves not only in the update of the partition of costumers into tables but also in sampling the partition of tables into dishes. See the Appendix D.1 for more details.

**Sampling $\gamma^{(L)}$.** To sample from the conditional posterior of $\gamma^{(L)}$, we need to take into account the way in which it affects the partition $\gamma^{(H)}$: $\gamma^{(L)}$ defines with its clusters the division of observations into restaurants; when the restaurants structure is changed, tables need to be rearranged and so do the dishes associated with them. Moreover for each value of $\gamma^{(H)}$, the partition of groups $\gamma^{(L)}$ is uniquely identified, meaning that the conditional posterior $p(\gamma^{(L)} | \gamma^{(H)}, \mathbf{x})$ is a point mass. Thus we cannot simply perform a Gibbs sampling step: a more elaborate procedure, with the flavor of a reversible-jump MCMC step, needs to be used.

The complexity of the structure of $\gamma^{(H)}$ prevents us from directly applying the reversible-jump MCMC framework (Green, 1995); instead we use a Metropolis-Hasting step, where the proposal updates the value of both partitions: we first propose a move for $\gamma^{(L)}$ using a split-merge move, and conditional on that move we propose a new value of $\gamma^{(H)}$, by rearranging

the tables in the new restaurant configuration. In split moves, tables are also split using a deterministic algorithm; in merge moves, tables are merged according to a sampling scheme that makes the split move reversible. See Appendix D.1.2 for more details.

### 5.3.3. Heuristic for posterior sampling

The combinatorial complexity of the space of partitions the MCMC needs to explore makes mixing difficult and times to convergence very long. To help with the exploration of the space, we devised a heuristic for posterior sampling. Instead of simultaneously sampling the LR and the HR partitions, we work iteratively:

1. Fixing $\gamma^{(L)}$ to the partition with $n$ clusters, initialize $\gamma^{(H)}$ to the partition with $n$ clusters and run the MCMC to sample only the HR partition and find an estimate $\widehat{\gamma^{(H)}}$.

2. Initialize $\gamma^{(L)}$ to the partition with $n$ clusters and $\gamma^{(H)}$ to $\widehat{\gamma^{(H)}}$. Run the standard MCMC algorithm, which samples both $\gamma^{(L)}$ and $\gamma^{(H)}$, and obtain an estimate of the LR partition $\widehat{\gamma^{(L)}}$.

3. Fixing $\gamma^{(L)}$ to $\widehat{\gamma^{(L)}}$, initialize $\gamma^{(H)}$ to the partition with $n$ clusters and run the MCMC to sample only the HR partition.

This algorithm allows us to have approximate samples from $\gamma^{(L)}$, from the second step, and from $\gamma^{(H)}$, from the third step. The estimates of the partitions can be found by minimizing the posterior expected variation of information (VI) distance (Wade et al., 2018). The choice of initializing the partitions to the "$n$-cluster" configuration is purely a practical one, because it reduces the change of the partition getting stuck in the configuration with one cluster, from which is often difficult to escape.

In the next section we compare its behavior with the standard MCMC algorithm, and the more restrictive HDP, on synthetic data analyses and real data.

## 5.4. Application to Areal Urban Data in Philadelphia

Areal data, i.e. data collected or aggregated within a region or areal unit, is present in many disciplines, such as economics, demography, or epidemiology. In fact, often for privacy reasons the precise geographic location is not collected or released, and the only available geographic information is the areal units the data "belong" to. The choice of the areal units can vary and, as a consequence, the data is often available at different granularity levels, or at multiple ones. For example, when the precise geographic information is known to the researchers but cannot be released, it can be aggregated at multiple possible levels, since different predictor information may be available for each level. Alternatively, data that is available at fine granularities can be aggregated to coarser granularities, to reduce the dimensionality of the problem. Examples of areal data are socio-economic indicators, crime levels or patients affected by a disease. All of these measures can be aggregated within lower resolution regions, by summing or averaging, and lead to a simpler and less detailed description of the phenomenon of interest. Especially when the number of regions is large, we are interested in clustering those regions in order to group together those with a similar value of the the parameter of interest; this approach can aide exploratory data analysis, to create interpretable maps, or can simply be used in mixture modeling to share information between different areas.

However, when the data are available at multiple granularities, it is unclear which level should be used for the analysis. For examples, cities in the US are divided in *census tracts*, which are then divided into *block groups*, which are further divided into *blocks*. Often the granularity level is chosen in an arbitrary way. Instead of fixing the resolution at which to perform our analysis, we want to be able to model the data at multiple resolutions. By doing this, a possible outcome in our analysis could be the *a posteriori* assessment of whether one resolution is overall better than the other. However, a more likely scenario seems to be one where one resolution is more appropriate for certain parts of the city and the other resolution is better for other urban areas. For example, central areas are more

densely populated and heterogeneous, compared to suburban areas. In the former we might expect a phenomenon like crime to be changing at a finer granularity than in the latter. In such case, analyzing the data at one level of resolution would be restrictive and important informations about the phenomenon could be lost. Instead, by modeling multiple resolution, we are able assess which level is more appropriate for a subset of the data.

In the criminology community there has been a particular interest in studying different levels of resolutions. In particular, studies have explored which granularity level had the largest spatial variation of crime and they have shown that the majority happens at micro levels such as street segments, rather than macro levels such as neighborhoods (see e.g. Schnell et al., 2017; Steenbeek and Weisburd, 2016; OBrien and Winship, 2017). However it is unclear if this finding is constant in different areas, such as central and suburban areas, or if some differences can be found. The maps of the spatial distribution of crime in Chicago reported in Schnell et al. (2017) suggest that our conjecture could prove to be true.

To cluster data at multiple resolutions, we can use nested models. While nested models have been studied for a long time, to the best of our knowledge they have never been used for multi-granularity modeling and clustering. In Rodríguez et al. (2008) for example they have been used for clustering hospital patients and the hospitals themselves based on their distribution of mixing components. While hospitals are supersets of patients, this cannot be considered as an example of multi-granularity data, as the interest lies in the measuring patient-level phenomena. Instead these models can be used to their full potential to describe situations when there is no specific preference for a granularity level.

Consider now a low resolution or coarse granularity level, $\{A_1, A_2, \ldots, A_m\}$ and a high resolution or fine granularity level $\{A_{11}, \ldots, A_{n_1 1}, \ldots, A_{1m}, \ldots, A_{n_m m}\}$. The high resolution areal unit $A_{lj}$ is fully contained (or nested) within the low resolution areal unit $A_j$ ($A_{lj} \subseteq A_j$). With a similar notation we denote with $y_{lj}$ the observation for the high resolution unit $A_{lj}$ and with $y_j$ the observation for the low resolution unit $A_j$; in situations where it is not available, $y_j$ will be artificially constructed by averaging (or summing) the observation

values of the units nested in $A_j$: $y_j = \frac{1}{n_j} \sum_{l=1}^{n_j} y_{lj}$.

Our analysis can be performed with only one level of high resolution of data, or with multiple levels of data, one for each resolution level. In the first case, low resolution units $A_j$ are clustered based on the proportion of mixture components $\mu_{lj}$ of their high resolution units; in the second case the low resolution partition is influenced by both the distribution of high resolution mixture components and by the low resolution data itself $y_j$. In either case, the high resolution clustering is influenced by the high resolution data $y_{lj}$ and by the partition of low resolution units, as it defines its hierarchical structure.

We first consider a model with only one level of high resolution of data:

$$\begin{aligned} y_{lj}|\mu_{lj} &\sim N(\mu_{lj}, \sigma^2) \\ \mu_{lj}|G_j &\sim G_j \\ G_j|Q &\sim Q \\ Q &\sim nHDP(\alpha_0, \alpha, \beta, H) \\ H &= N(0, {k_0}^{-1}\sigma^2) \\ \sigma^2 &\sim \text{Inv} - \text{Gamma}(\alpha_0, \beta_0). \end{aligned} \tag{5.1}$$

Note that $\sigma^2$ represents the within-cluster variance, because all the observations $y_{lj}$ corresponding to cluster $C_k = \{(l, j) : \mu_{lj} = \mu_k^*\}$ have the same mean $\theta_k^*$ and within-cluster variance $\sigma^2$. The cluster specific means instead are normally distributed with mean zero and variance $k_0^{-1}\sigma^2$, and it is important that this distribution covers the range of the data. Thus it is important to tune the prior for $\sigma^2$ so that we can specify the level of variation of the data within a cluster and to correctly choose $k_0$ so that the between-cluster variation covers the data. The hyper-parameters of the nested Hierarchical Dirichlet Process instead can be interpreted in the context of partitions of costumers into tables and of tables into dishes (for the high resolution partition) and of the partition of low resolution units.

Note that we can also describe this model in terms of the partitions induced by the nHDP: let $\gamma^{(L)}$ be the low resolution partition and $\gamma^{(H)}$ the high resolution one. Moreover, let $z_{jl}$ be the cluster membership of $y_{jl}$ induced by $\gamma^{(H)}$.

$$
\begin{aligned}
y_{lj}|z_{jl}, \boldsymbol{\mu}^*, \sigma^2 &\sim N(\mu^*_{z_{jl}}, \sigma^2) \\
\mu^*_k|\sigma^2 &\sim N(0, {k_0}^{-1}\sigma^2) \\
\gamma^{(H)}|\gamma^{(L)}, \alpha, \alpha_0 &\sim CRF(\alpha, \alpha_0; \gamma^{(L)}) \\
\gamma^{(L)}|\beta &\sim CRP(\beta) \\
\sigma^2 &\sim \text{Inv} - \text{Gamma}(\alpha_0, \beta_0).
\end{aligned}
$$

We then consider a model with two levels of data, in which we additionally model the low resolution data as normal $y_j|\theta_j \sim N(\theta_j, \sigma^2)$. The latent parameters $\theta_j$ are shared if two low resolution areas belong in the same cluster. More precisely, let $z_j$ be the low resolution cluster membership induced by $\gamma^{(L)}$, with $G_j = G^*_{z_j}$. Let $\theta^*_k$ be the unique values for the $\theta_j$'s and $y_j|\theta^*_k, z_j \sim N(\theta^*_{z_j}, \sigma^2)$. In other words, the low resolution partition induced by the $G_j$ affects the values of the $\theta_j$.

$$
\begin{aligned}
y_{lj}|z_{jl}, \boldsymbol{\mu}^*, \sigma_H^2 &\sim N(\mu^*_{z_{jl}}, \sigma_H^2) \\
\mu_k^*|\sigma_H^2 &\sim N(0, {k_0}^{-1}\sigma_H^2) \\
y_j|\boldsymbol{\theta}^*, z_j, \sigma_L^2 &\sim N(\theta^*_{z_j}, \sigma_L^2) \\
\theta_k^*|\sigma_L^2 &\sim N(0, k_1^{-1}\sigma_L^2) \\
\gamma^{(H)}|\gamma^{(L)}, \alpha, \alpha_0 &\sim CRF(\alpha, \alpha_0; \gamma^{(L)}) \\
\gamma^{(L)}|\beta &\sim CRP(\beta) \\
\sigma_H^2 &\sim \text{Inv} - \text{Gamma}(\alpha_H, \beta_H) \\
\sigma_L^2 &\sim \text{Inv} - \text{Gamma}(\alpha_L, \beta_L).
\end{aligned}
\tag{5.2}
$$

Given these two models for single and multi-level data, in the following section we describe the simulation analyses that compare the nHDP with other methods.

### 5.4.1. Synthetic analyses

In the following we describe different analyses of our method with some competitors. We first describe the analysis of a synthetic data set constructed on a subset of Philadelphia's block groups and census tracts. We then study the performance over a set of 50 synthetic datasets, in a simulation analysis of single-level data first and multi-level data subsequently.

**Synthetic data analysis** In the first analysis we test the correct behavior of our model and MCMC algorithm on a synthetic dataset. We generated partitions and data for a subset of the Philadelphia census tracts and block groups, and visually test the recovery of the 'true' partitions. In figure 22 we represent the data and partitions used to generate the data. After having designed the LR partition, we have chosen different high resolution cluster proportions for each low resolution cluster. For example, the blue cluster (labelled '2' in the left panel of figure 22) there is a 50% proportion of green and pink HR clusters (labeled '1' and '4' in the central panel of figure 22). In the light blue LR cluster (labelled

'1' in the left panel) there is a 100% proportion of the green HR cluster.

We ran our MCMC algorithm with the synthetic data represented in right panel of figure 22 and we considered two chains of 5000 iterations, discarding the first 2000 as burn in, and combined the two sets of samples. The hyper-parameters set were: $\eta_{LR} = \eta_{CT} = \eta_{TD} = 1$, $\alpha_0 = 3$, $\beta_0 = 1$, $k_0 = 0.01$. In figure 23 we report the estimated partitions. We can notice that the HR partition coincides with the true HR partition (used to generate the data). In the LR partition we notice a few discrepancies, in particular two census tracts are being misclassified and the total number of clusters estimated is less than the original total number of clusters, by one.

We also compare our model with using a simple HDP model, which even though it does not allow for multi-resolution modeling, it can be useful for estimating the high resolution partition. We use the HDP with two configurations: the first, represented in the top panels of figure 24, assumes the presence of one group of data, and fits the HDP with one restaurant; the second assumes the presence of as many groups as census tracts, and fits the HDP with a restaurant for each census tract. As we see in figure 24, the partitions recovered are very similar to the one used to generate the data. In the top panels we notice that it recovers a smaller number of clusters, while in bottom panel we notice that the number of clusters is larger than in the true partition.

**Simulation setting 1 (single-level data)** In a second analysis, we test how our method compares on a wider range of data and partitions, generated according to a finite dimensional approximation[1] of model 5.1: we sample the LR partition $\gamma^{(L)}$ from the DP($\eta_{LR}$); given the restaurant structure defined by $\gamma^{(L)}$, sample the HR partition $\gamma^{(H)}$ from the HDP($\eta_{CT}, \eta_{TD}$). Given $\gamma^{(H)}$, we consider equally spaced and centered around zero cluster-specific means $\mu_k$ and sample $y_i \sim N(\mu_{z_i}, \sigma^2)$. We specify these parameters to have somewhat separated

[1]We sample a finite dimensional approximation realization of a DP and HDP, using truncated stick-breaking, and then sample the categorical variables corresponding to the cluster memberships from the "finite dimensional" probabilities.

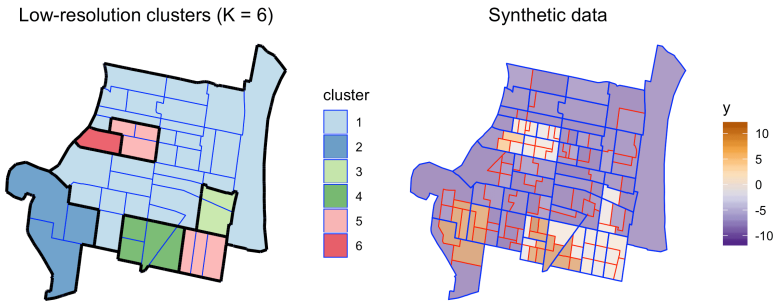


Figure 22: Synthetic data and corresponding partitions. Left panel: low-resolution partition, where census tracts (delineated in blue) in the same cluster have homogeneous proportions of HR clusters. Right panel: observed data for each block group (delineated in red).

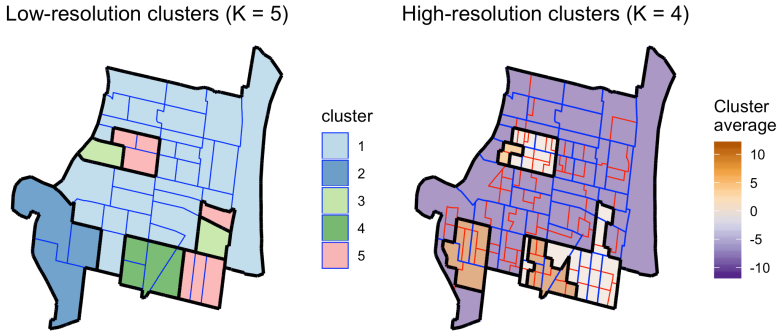


Figure 23: Partitions estimated using nHDP. Left panel: estimated low-resolution partition, with colors representing unique clusters. Right panel: estimated high-resolution partition, with colors representing the cluster average of $y$. This panel can be useful for noticing the similarity between the real data and the estimated cluster means.

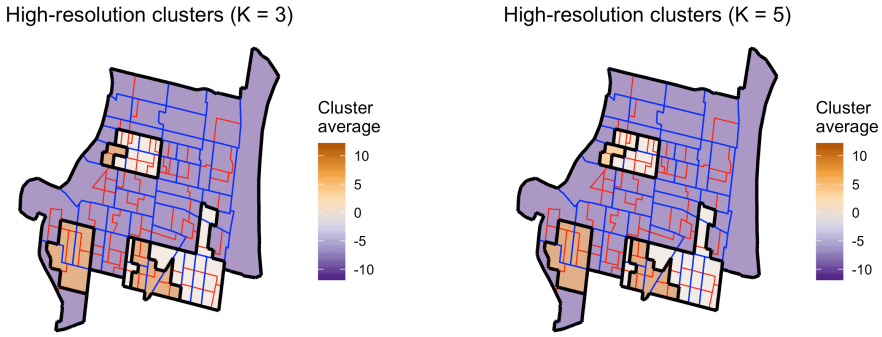


Figure 24: Partitions estimated using HDP with different number of groups (or restaurants). Left panel: estimated HR partition using one group (all the census tracts represent one restaurant); right panel: estimated HR partition using many groups, specifically one for each census tract. Colors represent the cluster average of $y$. LR partitions are not reported as the HDP does not estimate it, but it rather fixes it.

clusters that can be distinguished but not in a trivial way[2] We choose $\sigma = 0.5$ and $\mu_{k+1} - \mu_k = 2.5$. Moreover, to make the low-resolution clusters distinguishable, we make sure that the proportion of high-resolution clusters in each low-resolution clusters is actually different, by artificially merging low-resolution clusters that have the same proportions. We generated 50 partitions using $\eta_{LR} = \eta_{CT} = \eta_{TD} = 1$. We ran two MCMC chains for 10 thousands iterations each and discarded the first 2000 as burn-in.

| | $R^2_{\text{VI}}$ | $R^2_{\text{bayes}}$ | $\text{MSE}_{\text{VI}}$ | $\text{MSE}_{\text{bayes}}$ | $\text{VI}_{\text{HR}}$ | $\text{VI}_{\text{LR}}$ |
|---|---|---|---|---|---|---|
| nHDP | 0.739 | 0.813 | 0.509 | 0.356 | 0.324 | 1.058 |
| nHDP-heuristic | 0.782 | 0.849 | 0.101 | 0.093 | 0.169 | 0.000 |
| HDP-oracle | 0.792 | 0.845 | 0.059 | 0.049 | 0.097 | 1.050 |
| HDP-one | 0.760 | 0.841 | 0.232 | 0.148 | 0.289 | 1.116 |
| HDP-many | 0.795 | 0.849 | 0.067 | 0.054 | 0.135 | 4.344 |

Table 4: Results for simulation setting 1, with data generated according to the nHDP mixture model 5.1. The measures reported are averaged over the 50 simulated datasets: $R^2$ and MSE of the cluster-specific means, computed with the estimator given by partition minimizing VI (VI) and with the Bayesian posterior mean (bayes), VI distance between the estimated and true HR and LR partitions.

In table 4 we compare the performance of different models: the three models previously described ("nHDP", HDP with one restaurant "HDP-one", HDP with many restaurants "HDP-many"), the heuristic for the nHDP "nHDP-heuristic", and what can be considered as an oracle HDP, "HDP-oracle", in which the restaurant structure is defined by the 'true' low resolution partition. We report several measures. Firstly, we report an equivalent of $R^2$: denoting with $(\hat{\mu}_i)$ the estimate of the means, we measure with $R^2$ the ratio of the explained sum of squares and of the total sum of squares,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\mu}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

Secondly, we report the MSE for the estimation of the cluster specific means: $\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2$. Both these measures are reported for two estimators. The "VI" estimator finds an estimate $\hat{\gamma}_{\text{VI}}$ of the partition by minimizing the posterior expected variation of information

[2]We experimented also with settings where the clusters are well separated and we found that in such cases the performance is better.

(VI) distance (Wade et al., 2018), and computes $\hat{\mu}_i$ as the posterior mean conditional on $\hat{\gamma}_{\text{VI}}$. The "bayes" estimator averages over the partitions and estimates $\mu$ using the posterior mean. Thirdly, we report the VI distance between the estimated partition $\hat{\gamma}_{\text{VI}}$ and the true partition used to generate the data, for both the high-resolution and the low-resolution partition. When the HDP is used and as a consequence no LR partition is estimated ("HDPone","HDPmany" and "HDPoracle"), we used respectively the one-cluster partition, the one with the maximum number of clusters and the "true" low resolution partition. While we are not really interested in the VI distance between these artificial partition and the true one, we report it as a reference point.

We can see in table 4 that nHDP has the best estimation for the LR partition, as expected, and that "HDP-oracle" also almost the best performance overall. However, we can notice that while the HDP with one restaurant ("HDPone") has worse performance than nHDP, the HDP with many restaurants ("HDPmany") has performances comparable to the oracle HDP. We note that while it does not achieve the best performances, the heuristic for nHDP achieves a strong improvement with results slightly worse than HDP-oracle.

From these simulation analysis it seems that, when the interest lies in estimation of the high-resolution parameters, the nHDP does not have strong advantages over using a standard HDP ("HDP-many"), and neither does knowing the true low-resolution partition as in "HDP-oracle". However, it is important to remember that the purpose of the nHDP is not only better estimation of high resolution data, but also (and most importantly) simultaneous estimation of multiple resolution data, which is not achievable using the HDP.

**Simulation setting 2 (multi-level data)** In a third analysis we test the performance of our method under the multilevel data model 5.2. We use the same simulated datasets used for table 4, and construct the low resolution data by setting the mean values separated by five standard deviations, and sampling from independent normal distributions. In table 5 we compare the performance on high-resolution data for the same models reported in table 4, and we compare the performance on the low-resolution data for only the nHDP and the

heuristic for the nHDP.

| high-resolution | $R^2_{\text{VI}}$ | $R^2_{\text{bayes}}$ | $\text{MSE}_{\text{VI}}$ | $\text{MSE}_{\text{bayes}}$ | VI |
|---|---|---|---|---|---|
| nHDP | 0.785 | 0.840 | 0.097 | 0.094 | 0.128 |
| nHDP-heuristic | 0.791 | 0.847 | 0.052 | 0.049 | 0.099 |
| nHDP-oracle | 0.792 | 0.845 | 0.059 | 0.049 | 0.097 |
| HDP-one | 0.760 | 0.841 | 0.232 | 0.148 | 0.289 |
| HDP-many | 0.795 | 0.849 | 0.067 | 0.054 | 0.135 |
| low-resolution | $R^2_{\text{VI}}$ | $R^2_{\text{bayes}}$ | $\text{MSE}_{\text{VI}}$ | $\text{MSE}_{\text{bayes}}$ | VI |
| nHDP | 0.822 | 0.863 | 0.080 | 0.070 | 0.143 |
| nHDP-heuristic | 0.815 | 0.870 | 0.088 | 0.064 | 0.136 |

Table 5: Results for simulation setting 2, with data generated according to the nHDP mixture model 5.2. The measures reported are averaged over the 50 simulated datasets: $R^2$ and MSE of the cluster-specific means, computed with the estimator given by partition minimizing VI (VI) and with the Bayesian posterior mean (bayes), and VI distance between the estimated and true partitions. Measures are reported for both the high-resolution and the low-resolution data.

We notice how using the low resolution data helps the recovery of the high resolution structure too and helps the nHDP and nHDP-heuristic achieve much better performances compared to table 4. In particular, the "nHDP-heuristic" achieves as good performance as "HDP-oracle", and both show better results compared to the "HDP-many".

### 5.4.2. Clustering of crime in Philadelphia

The correct modeling of crime trends at a high resolution neighborhood level can benefit police departments, urban planner and city officials. We consider crimes aggregated within neighborhoods to recover areas of high and low crime levels; moreover, by considering neighborhoods measures we can incorporate our estimates with predictor information, which is only available at the neighborhood level.

We are interested in finding clusters of neighborhoods that display similar crime behaviors. The crime data are made publicly available by the Philadelphia Police department, which reports each reported crime type, GPS location and time. We focus on violent crimes, which consist of homicides, rapes, robberies and aggravated assaults, according to the definition by the Uniform Crime Reporting program of the FBI. We then aggregate these data by

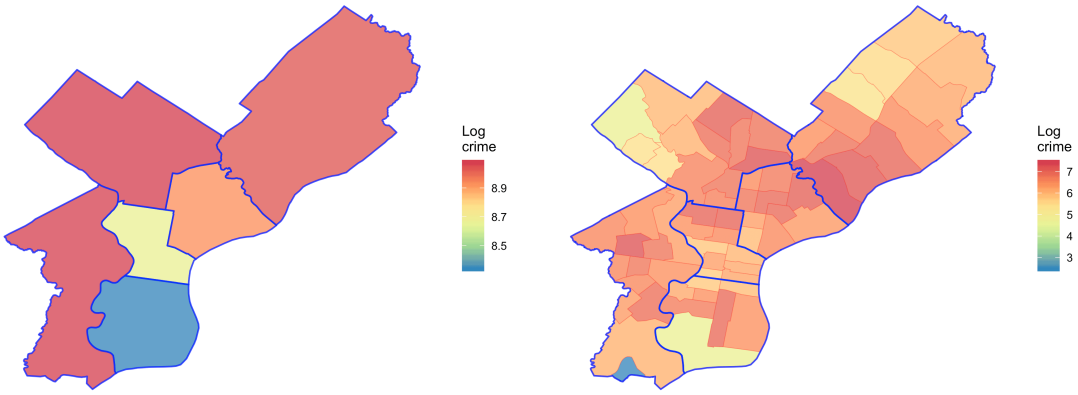


Figure 25: Maps of Philadelphia. Left panel: Philadelphia divided into 6 police divisions (PDV), delineated in blue. Right panel: Philadelphia divided into 66 police service areas (PSA). Each area is colored according to the mean level of crime in that region. Specifically, it represents the mean over 2006-2018 of the log-transformed crime counts.

counting how many violent crimes happened in each year within an areal unit boundaries and average those counts over the years from 2006 to 2018. Instead of averaging directly the yearly counts, we consider a log-transformation of the counts, to model the data as normally distributed. However, since some neighborhoods might have no crimes in some years, we consider the inverse hyperbolic-sine transformation, which is a good approximation to the logarithm but it's well defined at zero: if $c_t$ represents the count in year $t$ in an areal unit, then $y_t = \log(c_t + \sqrt{c_t^2 + 1}) - \log(2)$ is the transformed value. Thus our observables are the yearly average of the transformed number of crimes: $y = \frac{1}{T}\sum_{t=1}^{T} y_t$. Because of the approximate logarithmic transformation, note that the data at the lower resolution is not the sum or aggregation of the observations of the high-resolution units.

Several multi-resolution subdivisions of Philadelphia exist and in this work we consider two different ones: we analyze crimes aggregated at the block group and census tract levels, and we consider the subdivision induced by police departments.

**Police Service Areas and Police Divisions**

We first consider the subdivision of the City of Philadelphia defined by police divisions, districts and service areas. Philadelphia is divided in six police divisions (PDV), which are aggregations of police districts; there are twenty one police districts (PDS), each is under the leadership of a police Captain and is divided into two to four police service areas (PSA); Philadelphia is divided in 66 PSA, which are the smallest police units, each with their own team of police officers. To study heterogeneity of crime in police areal units we can consider the average of log-transformed yearly number of crimes within each PSA and within each PDV, the lowest and highest resolution levels. In Figure 25 we represent the real data, together with the boundaries of PSA (in red) and PDV (in blue). As the figure shows, there is some heterogeneity in the mean crime level in Philadelphia, and we can find some pockets of lower crime, together with one PSA of extremely low crime (the blue region in south west Philadelphia corresponds to the area of the airport).

For the analysis of the PSA data, we set the nHDP hyper-parameters to allow a relatively large number of high-resolution clusters: $\eta_{LR} = 5$, $\eta_{CT} = 4$ and $\eta_{TD} = 3$. We choose a rather informative prior for $\sigma_H^2$ ($\alpha_H = 160$ and $\beta_H = 10$, so that the prior median value for $\sigma_H$ is 0.25) to recover partitions with small within-cluster variance. For the model with multi-level data, model 5.2, we set a similarly informative prior on $\sigma_L^2$ ($\alpha_L = 10$ and $\beta_L = 0.1$, so that the prior median value for $\sigma_L$ is 0.1).

In Figure 26 and 27 we report the partitions recovered under these hyper-parameters, under model 5.1 and 5.2. The left panel refers to the low-resolution partition, while the right panel to the high-resolution one. All the partitions are estimated using posterior samples and minimizing the posterior VI distance (Wade et al., 2017).

Under both models, the high-resolution partition recovers several clusters, as desired. In particular, the regions of low crime in South Philadelphia and North West Philadelphia (labelled by 'A', 'B1' and 'B2') are separated from the rest, which is divided into areas of

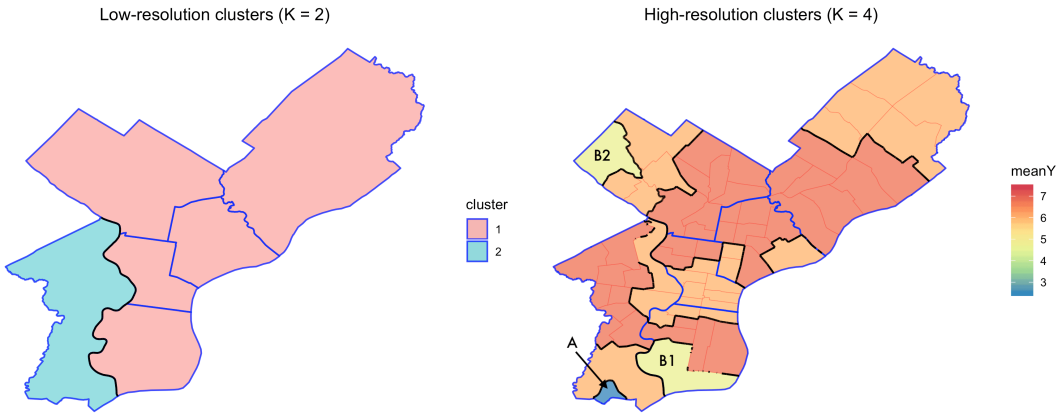


Figure 26: Partition recovered using an informative prior on $\sigma^2$. Top-left panel: the low resolution partition estimate, pictured using a different color for each cluster. Top-right panel: the high resolution partition estimate, pictured using the same color scale as the real data, i.e. the color for each cluster represent the mean of the data in such cluster.

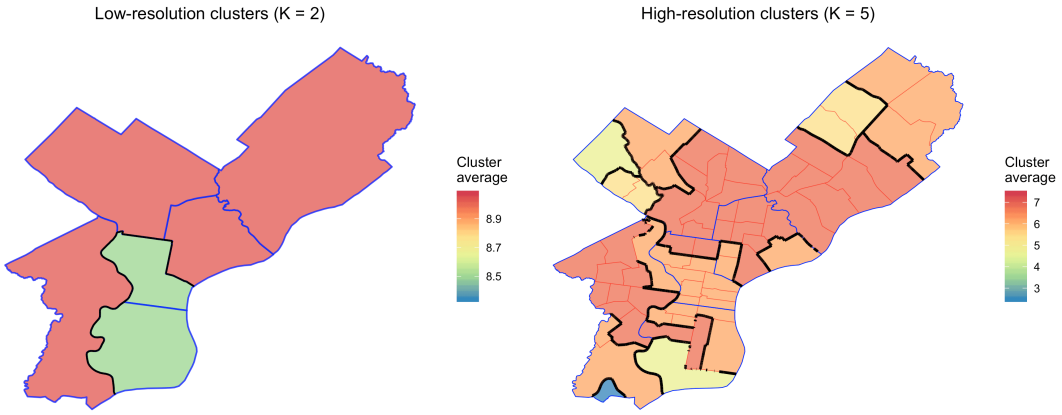


Figure 27: Partition recovered minimizing posterior VI, using non-informative prior on $\sigma^2_{LR}$ and an informative prior on $\sigma^2_{HR}$. Top-left panel: the low resolution partition estimate. Top-right panel: the high resolution partition estimate. Both maps use the same color scale as the real data, i.e. the color for each cluster represent the mean of the data in such cluster.

medium and high crime. We can instead notice some differences under the two models in the low-resolution partitions. Under model 5.1, depicted in Figure 26, the low-resolution partition is driven only by the proportion of high-resolution clusters. Since the western division (colored in blue in the left panel of Figure 26) is the only one containing low levels of crimes such as the airport area (labelled by 'A'), it gets separated in its own cluster. But under model 5.2, the low-resolution data drives the partition as well, and the partition recovered finds again two clusters, but this time separates the southern and central division (colored in green in the left panel of Figure 27) from the rest of the city.

While the analysis of Police Service Areas and Police Divisions can be considered instructive for the insights we have from using single or multi-level data, the disparity of resolution between Police Divisions and Police Service Areas makes this analysis not very useful in practice. In fact, the very high level of aggregation that we see in the Police Division data washes out any differences between Police Service Areas. This is a common problem, and it's known as the ecological fallacy. In the section, we will see how, with more similar granularity levels, the low-resolution data can be used to gain information about the high-resolution data.

**Census tracts and block groups: West Philadelphia**

We consider now the subdivision in block groups and census tracts, and we focus our analysis on the neighborhood of West Philadelphia. This part of the city, home of the University of Pennsylvania, is characterized by high heterogeneity in crime levels. For example, the number of violent crimes in each block group, averaged from 2006 to 2018, varies from a minimum of 6.2 (in a block group at 42th St and Baltimore St) to a maximum of 108.6 (in a block group at 46th St and Market St), two parts of the city that dist approximately 1 mile. Similarly, the minimum and maximum average number of violent crime at the tract level happen at a distance of approximately 1.5 miles (even though they correspond to different parts of West Philadelphia).

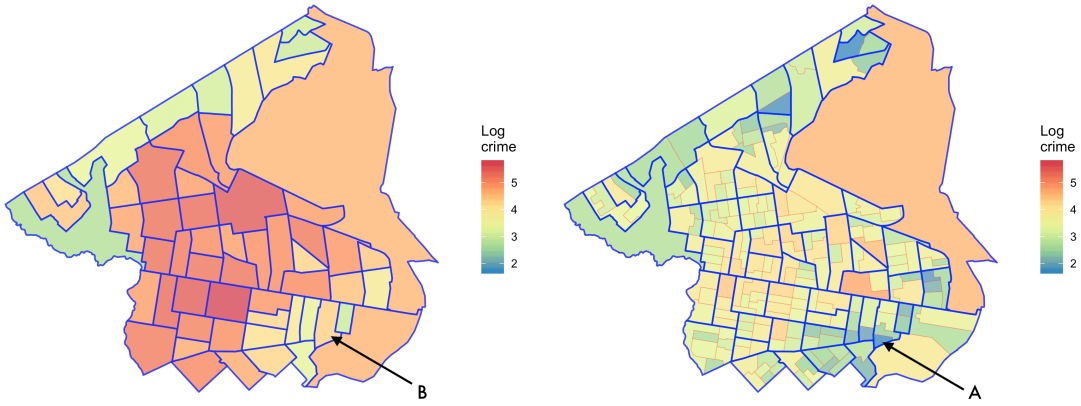


Figure 28: Log-transformed number of violent crimes, averaged over the years from 2006 to 2018, aggregated at the census tract (left panel) and block group resolutions (right panel).

In fact, when working with different resolutions things can actually look really different depending on the level chosen. Instead of limiting ourselves to only working with census tracts or only block groups, we model both. We consider the log-transformed number of violent crimes in a unit, averaged over the years from 2006 to 2018.

In figure 28 we represent the data. It is interesting to note that the block group with the smallest average number of crime (labelled with A) is contained in a census tract (labelled with B) whose crime level is not among the lowest ones.

We analyzed this region with the multi-data model 5.2, after rescaling the two variables $y_{lj}$ and $y_j$ to be centered at zero and have unit variance.

In figure 29 we report the partitions recovered when we chose highly informative hyper-parameters that would allow for relatively small variation within clusters. We notice that various regions in the north of West Philadelphia seem to be described by the clusters found at the lower resolution of census tracts, while regions in the center and south-east seem to be better described by the higher resolution of block groups. For example, the region in the center, highlighted by a circle in the right panels, displays a much finer partition at the block group resolution in the bottom right panel, which is not well described by the low resolution clustering in the top right panel. While the data aggregated at the census

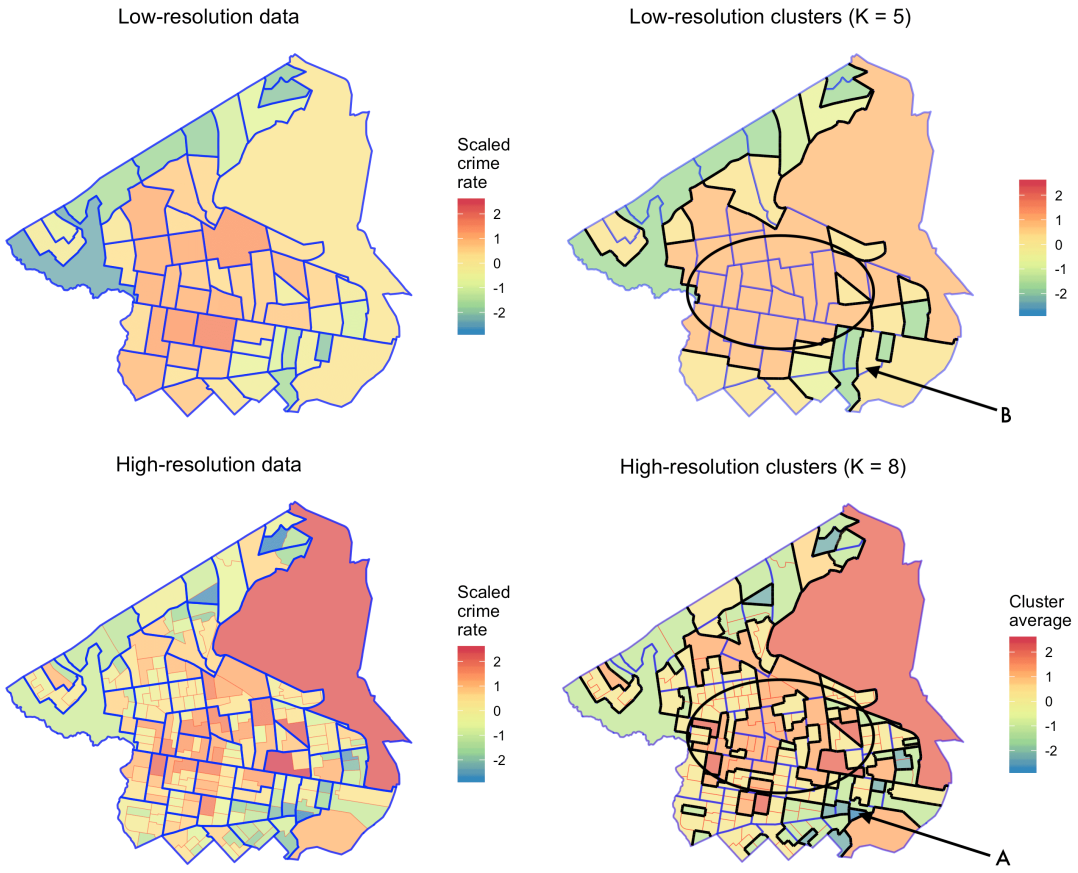


Figure 29: Real data centered and scaled (left panels) and estimated partitions (right panels) for census tracts data (top panels) and block groups data (bottom panels).

tracts level identifies such area with moderately large levels of crime, colored in orange, the finer resolution data allows us to recognize many block groups with medium level of crime, colored in yellow, together with the higher crime level areas, colored in orange and red. Moreover, the region around the eastern part of Baltimore street, labelled A and B in respectively the block group and census tract plots, is characterized by areas with extremely low levels of crime, colored in blue, which are adjacent to areas with medium crime, colored in yellow. It is also interesting to notice that the clusters found at the higher resolutions are not nested in the lower resolution clusters. This flexibility is allowed by the nested Hierarchical Dirichlet Process but would not be possible under the nested Dirichlet Process.

## 5.5. Discussion

Choosing a level of resolution when working with spatial data is often challenging and can affect results. In this work we explored the multi-resolution approach, that instead of choosing one level, simultaneously models multiple resolutions. In particular we propose using nested models for clustering data at multiple resolutions and we extended the nested Dirichlet Process (Rodríguez et al., 2008) to generate more flexible partitions: the nested Hierarchical Dirichlet Process allows us to find partition of low resolution and high resolution data, while not imposing restricting constraints in the clusterings structure.

We proposed a split-merge MCMC algorithm to draw samples from our model and we show its performance in simulated analysis, in cases both where the focus is on mixture modeling of only high-resolution data, and where the interest lies in partitions of both low and high resolution data. We analyze crime levels in Philadelphia, aggregated within police districts boundaries and within census tracts and block groups.

While split-merge algorithms are known for having good mixing properties, in such high-dimensional space it might be convenient to develop a Gibbs sampling step for the updating of the low-resolution partition. While this cannot be simply computed analytically, it is

possible to numerically compute the conditional probabilities and it could reveal a more efficient way of sampling from the posterior distribution.

However, even the most efficient MCMC algorithm can be extremely slow for high dimensional problems such as this when the number of units or when the levels of resolutions increase. It is important to develop more efficient ways of analyzing multi-resolution data. It is possible to do this, under a different model, using Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) and it will be the object of future research.

While in this work we focused on the application to urban crime, many other domains can benefit from multi-resolution modeling, from the social sciences to epidemiology, which also use data aggregated within areal units. Another example of a domain that could benefit from multi-resolution models, while not using geographical areal data, is neuroimaging: by clustering both brain regions and voxels, multi-resolution models could allow the discovery of similarities between brain regions while recovering patterns at the voxel level.

# Chapter 6

# Conclusion and Future Directions

Spatial data often varies in a smooth, continuous manner, but is also characterized by occasional discontinuities, where near locations display starkly different behaviors. We explore models and methods to combine smoothness with discontinuities, in particular through partitions of areal data.

We consider the applied problem of studying the spatial variation of crime dynamics in the City of Philadelphia. Accurate modeling of urban crime dynamics can offer benefits to law enforcement officials for deployment of resources for public safety, to urban planners for better understanding of the effects of socio-economic factors and the built environment on crime, and to city officials for improvement of the quality of life in the city. We consider crime counts in local neighborhoods in the period of 2006-2015, combine them with socio-economic and built environment predictors and study the percentage linear change over time. We compare several Bayesian approaches while sharing information either globally or locally across the city. We find that imposing local shrinkage between proximal neighborhoods using a spatial conditional autoregressive (CAR) prior achieves the best out-of-sample predictions of violent crime. We also recover spatial discontinuities by allowing the adjacency matrix of the CAR model to vary. We find that discontinuities for the trend over time and for the mean level of crime do not necessarily overlap.

We then extend this model, by incorporating the spatially smoothing CAR model with spatial clustering, to recover parts of the city displaying different behaviors and to allow for discontinuities by assuming a "CAR-within-cluster" model. We do not assume that the partitions in the mean level of crime and in the trends over time coincide; therefore we try to identify pairs of partitions of areal units. For such vast discrete spaces, stochastic search methods are computationally prohibitive when the number of units to partition is large.

We introduce a method that by solving a single optimization problem, approximates the posterior distribution by identifying partitions with highest posterior probability. Through synthetic data we show that our Particle Optimization method has good estimation and partition recovery properties. We also use it to analyze the Philadelphia crime data and find that many of the recovered discontinuities coincide with natural or built barriers, such as parks and streets.

In our analysis of crime in Philadelphia, we find that the prior can have a strong influence on the inferences. We compare some distributions for partitions of areal data and find that many are not mathematically manageable, because they induce non exchangeable partitions and they are specified up to a normalizing constant. We empirically study some properties and find that some distributions, even though designed with the purpose of adapting the Dirichlet Process prior to areal data, completely distort its behavior. Other distributions display more similar behaviors to the Dirichlet Process prior but pay the price of not inducing strong spatial behaviors. We propose two distributions, the SBM-PPMx and the Areal-PPMx. The first has nice mathematical properties of coherence across sample sizes, but does not show good empirical properties. The second has the opposite behavior: it does not satisfy the coherence property but works well empirically.

Finally, we extend the problem of clustering to multiple resolution data. Areal data can be aggregated at different granularities, and choosing one can be restrictive. It is possible to cluster data at multiple levels using nested models such as the nested Dirichlet Process, but this model induces specific restrictions on the partitions induced. To allow for more flexible partitions, we introduce the nested Hierarchical Dirichlet Process, and we describe a posterior sampling algorithm. We show the performance of our model on synthetic and real data and we find that sometimes the standard MCMC algorithm has difficulties exploring the complex discrete space. We propose a heuristic that approximates the sampling distribution and achieves performance almost as good as the oracle's. We apply our model to crime data. We find that overall higher variation is present at higher-resolutions, but that

suburban areas display less high-resolution heterogeneity, and thus can be appropriately described by lower-resolution levels.

## 6.1. Next Steps

In this thesis we have approached various challenges posed by clustering areal data, but more work needs to be done.

Our Particle Optimization method works very well when the model is conjugate and we can explicitly compute the marginal probability of the data given the unknown partitions. However, many more complex models don't have such feature. Neither do models using less mathematically convenient prior distributions for partitions, such as those described in this work. Being able to extend this method to more general problems would be certainly useful in many settings.

The problem of choice of prior distributions for areal data is far from being solved. The complex dependence induced by areal data and the combinatorial nature of the problem make it difficult to derive exact properties. In this work we show that by empirical exploration of these properties we can discover more about these distributions. More work needs to be done to construct an auxiliary model extending the PPMx framework to areal data, which displays the property of coherence across sample sizes and simultaneously has good empirical properties.

When clustering data at multiple resolutions, the computational aspect is even more important, and different algorithms can be considered to improve convergence and mixing. However, even the most efficient MCMC algorithm can be extremely slow for high dimensional problems such as these when the number of units or when the levels of resolutions increase. It is important to develop more efficient ways of analyzing multi-resolution data; for example, by adapting Bayesian Additive Regression Trees (BART) we could develop a model for multiple resolution data that can be computationally simpler.

# APPENDIX A

# Spatial Modeling of Trends in Crime over Time

## A.1. Maps of Data in Philadelphia

Figure 30 (left) gives a map outlining the 1336 block groups in Philadelphia. Figure 30 (right) shows population count for each block group in Philadelphia.

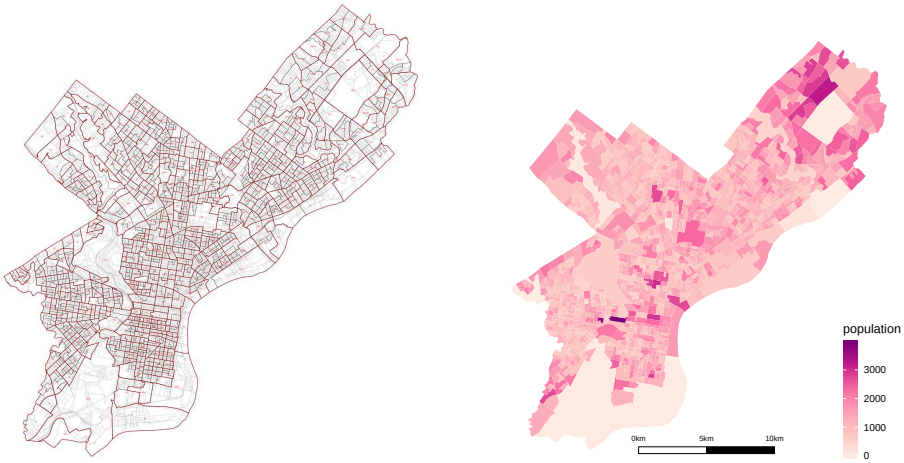


Figure 30: **Left:** Map of Philadelphia divided into census tracts (red lines) and block groups (black lines) by US Census Bureau. **Right:** Population count by block group in Philadelphia. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

In Figure 31 (left), we give the count of violent crimes per year in each block group, averaged over the years 2006-2015. We see substantial heterogeneity across block groups in the average counts of violent crimes per year. There are several outlying values: particular block groups that have much higher average violent crime counts. The largest among these is the Market East neighborhood in central Philadelphia.

These outlying neighborhoods motivate us to examine violent crime totals on the log scale. In Figure 31 (right), we give the average of the logarithm of the count of violent crimes per

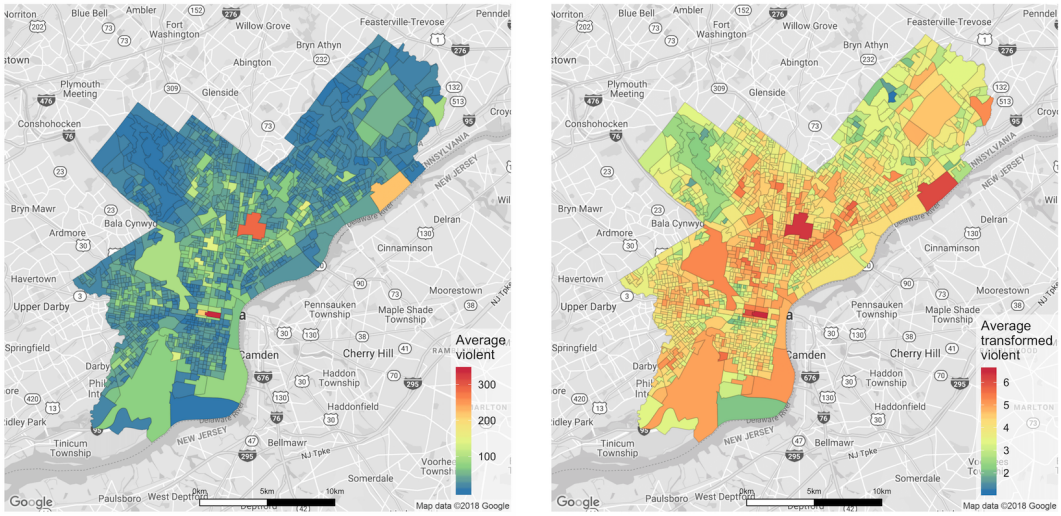


Figure 31: Distribution of violent crime over the block groups of Philadelphia. **Left:** violent crimes per block group, averaged over the years from 2006 to 2015. **Right:** logarithm of violent crimes per block group, averaged over the years from 2006 to 2015. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

year in each block group, averaged over the years 2006-2015. We can see more details of the spatial distribution of violent crime on the log scale. Modeling crime on the log scale has the additional benefit that changes in log crime can be interpreted as percentage changes in crime.

In both Figure 31 (left) and (right), we see evidence of spatial correlation in violent crime totals between proximal block groups throughout the city.

## A.2. Gibbs sampling

In section 2.3, we described the different models considered in this work; here we illustrate the Gibbs sampling strategy to sample from their posterior distributions. We are mainly interested in the coefficients $\boldsymbol{\gamma}$ for the predictors and in the collection of neighborhood-specific coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. $\boldsymbol{\gamma}$ have a Normal prior distribution with mean zero and covariance matrix proportional to the identity matrix. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ also have Normal prior distributions with mean zero, but have different covariance matrices depending on the model. In the global shrinkage model, the covariance matrices are proportional to the identity matrix, while in

the local shrinkage CAR model the covariance matrices depend on the Laplacian matrix of the geography.

We denote with $\mathbf{Y}$ the $N$-dimensional vector (where $N = nT$) found by concatenating the $y_{it}$'s, ordered by block group: $\mathbf{Y} = (y_{11}, \ldots, y_{1T}, \ldots, y_{n1}, \ldots, y_{nT})$; moreover let $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ represent the collection of the coefficients. Let the matrix $\mathbf{X}$ be designed so that the covariates corresponding to block group $i$ at time $t$ are contained in row $(i-1)T + t$ and by multiplying this row with the vector of parameters we get $X^{\gamma}_{(i-1)T+t}\boldsymbol{\theta} = \sum_j \gamma_j z_{ij} + \alpha_i + \beta_i t$. With this notation, the conditional distribution of the data is given by $\mathbf{Y}|\boldsymbol{\theta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I})$.

We can also combine the prior distributions of $\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ to get the distribution of $\boldsymbol{\theta}$: let $\boldsymbol{\theta}_0 = (\mathbf{0}, \alpha_0\mathbf{1}, \beta_0\mathbf{1})$ a $(d+2n)$-dimensional vector representing the conditional mean of $\boldsymbol{\theta}$ and let $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0^{-1}$ be the block matrix representing its precision matrix. Since $p(\boldsymbol{\gamma}) = \mathrm{N}(\mathbf{0}, \tau_\gamma^2 \cdot \mathbf{I})$, the first $d \times d$ diagonal block of $\boldsymbol{\Omega}_0$ is equal to $\tau_\gamma^{-2}\mathbf{I}$; the next two $n \times n$ diagonal blocks instead are the precision matrices of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$: $\tau_\alpha^{-2}\boldsymbol{\Sigma}^{-1}$ and $\tau_\beta^{-2}\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma} = \mathbf{I}$ in the global shrinkage model and $\boldsymbol{\Sigma}^{-1} = [\rho(\mathbf{D}_W - \mathbf{W}) + (1-\rho)\mathbf{I}]$ in the spatial CAR model; the remaing blocks are zero matrices. Then $p(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \tau_\alpha^2, \tau_\beta^2) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\boldsymbol{\Omega}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right)$. Moreover, we set non-informative flat priors on $\alpha_0, \beta_0$, and the variance hyper-parameters $\sigma^2$, $\tau_\gamma^2$, $\tau_\alpha^2$ and $\tau_\beta^2$ have Inverse-Gamma priors, tuned in an Empirical Bayes fashion.

**Posterior conditional distribution of $\boldsymbol{\theta}$** With this notation, we can find the conditional posterior distribution of $\boldsymbol{\theta}$ as we would do in a usual linear regression:

$$\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\theta}_0, \sigma^2, \boldsymbol{\tau}^2 \sim N(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$$

where

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left(\boldsymbol{\Omega}_0 + \mathbf{X}^T\mathbf{X}/\sigma^2\right)^{-1}\left(\boldsymbol{\Omega}_0\boldsymbol{\theta}_0 + \mathbf{X}^T\mathbf{Y}/\sigma^2\right)\\ \mathbf{V}_{\boldsymbol{\theta}} &= \left(\boldsymbol{\Omega}_0 + \mathbf{X}^T\mathbf{X}/\sigma^2\right)^{-1}.\end{aligned}$$

**Posterior conditional distribution of $\boldsymbol{\theta}_0$** Similarly, the posterior distribution of the mean hyper-parameters $\alpha_0$ and $\beta_0$ can be found as

$$\alpha_0|\boldsymbol{\alpha}, \tau_\alpha^2 \sim N\left(\frac{\mathbf{1}^T\Sigma_\alpha^{-1}\boldsymbol{\alpha}}{\mathbf{1}^T\Sigma_\alpha^{-1}\mathbf{1}}, \frac{\tau_\alpha^2}{\mathbf{1}^T\Sigma_\alpha^{-1}\mathbf{1}}\right)$$
$$\beta_0|\boldsymbol{\beta}, \tau_\beta^2 \sim N\left(\frac{\mathbf{1}^T\Sigma_\beta^{-1}\boldsymbol{\beta}}{\mathbf{1}^T\Sigma_\beta^{-1}\mathbf{1}}, \frac{\tau_\beta^2}{\mathbf{1}^T\Sigma_\beta^{-1}\mathbf{1}}\right).$$

**Posterior conditional distribution of $\sigma^2$, $\tau_\gamma^2$, $\tau_\alpha^2$ and $\tau_\beta^2$** For the variance hyper-parameters $\sigma^2, \tau_\gamma^2, \tau_\alpha^2$ and $\tau_\beta^2$, the prior distributions are

$$\sigma^2 \sim \text{Inv-Gamma}(a_\sigma, b_\sigma)$$
$$\tau_\alpha^2 \sim \text{Inv-Gamma}(a_\alpha, b_\alpha)$$
$$\tau_\beta^2 \sim \text{Inv-Gamma}(a_\beta, b_\beta)$$
$$\tau_\gamma^2 \sim \text{Inv-Gamma}(a_\gamma, b_\gamma).$$

where the hyper-parameters are tuned in an empirical Bayes fashion so that the prior mean of the variance parameters is equal to the variance estimated from the model with no shrinkage, and the prior variance is small. The conditional posterior distributions are also Inverse-Gamma:

$$\sigma^2|\mathbf{Y}, \boldsymbol{\theta} \sim \text{Inv-Gamma}\left(a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \mathbf{z}_i^T\boldsymbol{\gamma} - \alpha_i - t\beta_i)^2\right)$$
$$\tau_\gamma^2|\mathbf{Y}, \boldsymbol{\gamma} \sim \text{Inv-Gamma}\left(a_\gamma + d/2, b_\gamma + \frac{1}{2}\sum_{j=1}^{d}\gamma_j^2\right)$$
$$\tau_\alpha^2|\mathbf{Y}, \boldsymbol{\alpha}, \alpha_0 \sim \text{Inv-Gamma}\left(a_\alpha + \frac{n}{2}, b_\alpha + (\boldsymbol{\alpha} - \alpha_0\mathbf{1})^T\boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{\alpha} - \alpha_0\mathbf{1})/2\right)$$
$$\tau_\beta^2|\mathbf{Y}, \boldsymbol{\beta}, \beta_0 \sim \text{Inv-Gamma}\left(a_\beta + \frac{n}{2}, b_\beta + (\boldsymbol{\beta} - \beta_0\mathbf{1})^T\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta} - \beta_0\mathbf{1})/2\right).$$

**Posterior conditional distribution of $\rho$** The prior distribution on $\rho$ is Beta$(10, 10)$, and since its conditional posterior distribution does not have a closed form, we sample this

parameter with a Metropolis Hasting procedure. Given its past value $\rho^{t-1}$ we propose a new candidate $\rho^*$ with density $g(\rho^*|\rho^{t-1}) = \text{Beta}(b\rho^{t-1}/(1-\rho^{t-1}), b)$; this parametrization allows the mean to be $\rho^{t-1}$ and the variance to be small when we choose $b = 10$. The acceptance probability is then

$$a = 1 \wedge \frac{p(\rho^*|e.e.)}{p(\rho^{t-1}|e.e.)} \frac{g(\rho^{t-1}|\rho^*)}{g(\rho^*|\rho^{t-1})}$$

where the posterior conditional distribution $p(\rho|e.e.)$ is proportional to the product of the prior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ given $\rho$ and the prior of $\rho$. We use the notation $e.e.$ to denote "everything else", i.e. the current values of all other parameters in the model.

**Posterior conditional distribution of W** Finally, in model (2.13)-(2.14) we allow the adjacency matrix itself to be random. We consider all the pairs of regions that share a border ($w_{ij} = 1$) and we allow those borders to potentially become barriers ($w_{ij} = 0$). We model these variable weights as $w_{ij}^{\alpha}|\phi^{\alpha} \sim \text{Bern}(\phi^{\alpha})$ and independently, $w_{ij}^{\beta}|\phi^{\beta} \sim \text{Bern}(\phi^{\beta})$, with $\phi^{\alpha}, \phi^{\beta} \overset{iid}{\sim} \text{Beta}(1, 9)$.

With these prior distributions, the conditional posterior distribution for $\mathbf{W}^{\alpha}$ is

$$\begin{aligned}
p(\mathbf{W}^{\alpha}|e.e.) &\propto p(\boldsymbol{\alpha}|\alpha_0, \tau_{\alpha}^2, \rho, \mathbf{W}^{\alpha}) p(\mathbf{W}^{\alpha}|\phi^{\alpha}) \\
&\propto \det(\boldsymbol{\Sigma}_{\alpha}^{-1})^{1/2} \exp\left(-\frac{1}{2\tau_{\alpha}^2}(\boldsymbol{\alpha} - \alpha_0\mathbf{1})^T \boldsymbol{\Sigma}_{\alpha}^{-1}(\boldsymbol{\alpha} - \alpha_0\mathbf{1})\right) p(\mathbf{W}^{\alpha}|\phi^{\alpha}) \\
&\propto \det(\boldsymbol{\Sigma}_{\alpha}^{-1})^{1/2} \exp\left(-\frac{\rho}{2\tau_{\alpha}^2}(\boldsymbol{\alpha} - \alpha_0\mathbf{1})^T (\mathbf{D}_{W^{\alpha}} - \mathbf{W}^{\alpha})(\boldsymbol{\alpha} - \alpha_0\mathbf{1})\right) p(\mathbf{W}^{\alpha}|\phi^{\alpha}).
\end{aligned}$$

Note that, because of the determinant term, the entries of $\mathbf{W}^{\alpha}$ are not independent *a posteriori*. Thus we sample each entry $w_{ij}^{\alpha} = w_{ji}^{\alpha}$ conditional on the rest of the matrix $\mathbf{W}_{-ij}^{\alpha}$ as $p(w_{ij}^{\alpha} = 1|e.e.) = q$, where

$$\frac{q}{1-q} = \sqrt{\frac{\det(\boldsymbol{\Sigma}_{\alpha}^{-1}(w_{ij}^{\alpha} = 1))}{\det(\boldsymbol{\Sigma}_{\alpha}^{-1}(w_{ij}^{\alpha} = 0))}} \exp\left(-\frac{\rho}{2\tau_{\alpha}^2}(\alpha_i - \alpha_j)^2\right) \frac{\phi^{\alpha}}{1-\phi^{\alpha}}.$$

A highly similar procedure (with the obvious substitutions) is used to sample the entries of $\mathbf{W}^{\beta}$.

**Posterior conditional distribution of $\phi$** To express the prior information that only a small percentage of the borders should be turned into barrier, the prior distribution of $\phi^{\alpha}$ and $\phi^{\beta}$ is Beta$(1, 9)$. Since the $w_{ij}^{\alpha}$ and $w_{ij}^{\beta}$ are Bernoulli distributed, the posterior distributions for $\phi^{\alpha}$ and $\phi^{\beta}$ are

$$\phi^{\alpha}|\mathbf{W}^{\alpha} \sim \text{Beta}\left(1 + \sum_{(i,j)\in I} w_{ij}^{\alpha}, 9 + \sum_{(i,j)\in I} (1 - w_{ij}^{\alpha})\right)$$
$$\phi^{\beta}|\mathbf{W}^{\beta} \sim \text{Beta}\left(1 + \sum_{(i,j)\in I} w_{ij}^{\beta}, 9 + \sum_{(i,j)\in I} (1 - w_{ij}^{\beta})\right).$$

## A.3. MCMC Implementation Details

The results reported in Table 1 for the hierarchical models have been implemented using Gibbs sampling. In particular, for each model 1000 samples where used, after discarding a burn-in period of 50 iterations and thinning every 2 samples. By running multiple chains and superimposing their trace plots, we noted that the convergence happened after a relatively short time and that samples were not highly correlated.

## A.4. Prior Robustness for Variance Hyperparameters

In Section 2.3, we used priors for the variance parameters with hyper-parameters that were tuned in an Empirical Bayes fashion. In this section, we show that highly similar results are obtained when using more non-informative prior distributions on these variance parameters.

In particular, we consider a uniform prior on $\boldsymbol{\gamma}, \log\sigma, \tau_\alpha$ and $\tau_\beta$, which is equivalent to:

$$p(\boldsymbol{\gamma}) \propto 1$$
$$p(\sigma^2) \propto \sigma^{-2}$$
$$p(\tau_\alpha^2) \propto \tau_\alpha^{-1}$$
$$p(\tau_\beta^2) \propto \tau_\beta^{-1}.$$

Table 6 is equivalent to Table 1 but with results from the estimated models that use the non-informative priors given above. Only the global, spatial CAR and variable border model results are reported since priors are not involved in the no-shrinkage model (2.5).

| Model | $\text{MSE}_{\text{in}}$ | $\text{MSE}_{\text{out}}$ | $\text{MSE}_{\text{cv}}$ | Moran's $I$ |
|---|---|---|---|---|
| Separate $\alpha_i$ , $\beta_i$ Models | | | | |
| Global Shrinkage | 0.0698 | 0.1080 | 0.0927 | 0.17 |
| Spatial CAR | 0.0701 | 0.1052 | 0.0922 | 0.61 |
| Variable Borders | 0.0706 | 0.1069 | 0.0927 | 0.48 |

Table 6: Comparison of predictive accuracy between the different models outlined in Section 2.3 with non-informative priors on the hyper parameters. The mean squared error for both in-sample and out-of-sample predictions are provided. We also provide the Moran's $I$ measure of spatial correlation calculated on the estimated time trends $\beta_i$ from each model.

Examining Table 6, we see almost the exact same predictive results as the predictive results given in Table 1. There are very slight numerical differences in the Spatial CAR in-sample error, the Global Shrinkage out-of-sample error and the Variable Borders Moran's I, but these differences could easily be attributed to MCMC sampling variability.

## A.5. Additional Model Results

In Section A.5.1 we report the numerical estimates of the partial effects, which are shown in Figure 2. In Section A.5.2 we describe the results from a model with variable borders for $\alpha_i$ but fixed borders for $\beta_i$. In section A.5.3 we provide different visualizations of the variability in the estimated neighborhood-specific coefficients $\alpha_i$ and $\beta_i$.

### A.5.1. Numerical results for partial effects

In table 7 we report the numerical values corresponding to the partial effects shown in Figure 2. For each model with neighborhood-specific coefficients outlined in Section 2.3, we report maximum likelihood estimates, standard errors, posterior means and posterior standard deviations.

| | No shrinkage | | Global shrinkage | | Spatial CAR | | Variable borders | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | St.Error | Mean | SD | Mean | SD | Mean | SD |
| log.income | -0.186 | 0.009 | -0.186 | 0.022 | -0.113 | 0.023 | -0.099 | 0.021 |
| sqrt.poverty | 0.182 | 0.009 | 0.182 | 0.023 | 0.095 | 0.020 | 0.110 | 0.018 |
| segregation | 0.010 | 0.005 | 0.010 | 0.013 | -0.023 | 0.017 | -0.013 | 0.017 |
| sqrt.vacantprop | 0.116 | 0.006 | 0.115 | 0.014 | 0.052 | 0.015 | 0.045 | 0.014 |
| sqrt.comresprop | 0.227 | 0.005 | 0.223 | 0.013 | 0.241 | 0.011 | 0.240 | 0.011 |
| pop.total | 0.216 | 0.005 | 0.212 | 0.013 | 0.263 | 0.011 | 0.317 | 0.012 |

Table 7: Estimate and standard error for each coefficient $\gamma_j$. For the Bayesian models, we report the mean and the standard deviation from 1000 independent draws from the posterior distribution.

### A.5.2. Random borders for only $\boldsymbol{\alpha}$

Given the smaller number of barriers detected for the $\beta_i$'s compared to the $\alpha_i$'s in Figure 5, we also implemented an alternative model where the adjacency structure $\mathbf{W}^\beta$ for the $\beta_i$'s is considered fixed, and only the adjacency matrix $\mathbf{W}^\alpha$ for the $\alpha_i$'s is allowed to vary.

The in-sample MSE of 0.0711 for the model with variable $\mathbf{W}^\alpha$ and fixed $\mathbf{W}^\beta$ is worse then the in-sample MSE of 0.0706 for the model with variable $\mathbf{W}^\beta$ and $\mathbf{W}^\alpha$. However, the variable $\mathbf{W}^\alpha$ and fixed $\mathbf{W}^\beta$ model does have a slightly better out-of-sample MSE of 0.1050 compared to the out-of-sample MSE of 0.1069 for the model with variable $\mathbf{W}^\beta$ and $\mathbf{W}^\alpha$.

These results provide a further indication that there is stronger signal in the data for detecting discontinuities for the mean level of crime between neighborhoods compared to discontinuities for the trend in crime over time between neighborhoods. However, we still report the results for the model with variable $\mathbf{W}^\alpha$ and $\mathbf{W}^\beta$ in Section 2.4, as it provides additional insight and interpretation in Figure 4.

### A.5.3. Variability of neighborhood-specific coefficients

In Section 2.5.4, we report the neighborhoods with the highest and lowest estimates of the mean level of crime $\alpha_i$ and time trend $\beta_i$. As a supplement to these results, in Figure 32 we map the Philadelphia neighborhoods which are "significant", in the sense that their 95% credible intervals do not contain the global mean $\bar{\alpha}$ or $\bar{\beta}$ across all neighborhoods in the city.

In Figure 32, we see many more neighborhoods with significant differences in terms of their $\alpha_i$'s, which is another indication that the variation in the mean level of crime is larger than the variation in the time trend in crime ($\beta_i$'s). In the plot for $\beta_i$'s, we find a smaller number of neighborhoods with values that are significantly different than the overall mean, but the existence of these neighborhoods confirms the presence of the space-time interaction found in previous studies (Law et al., 2014; Li et al., 2014).

In Figure 33, we visualize the width of the 95% credible intervals for each neighborhood-specific $\alpha_i$ and $\beta_i$. It is interesting to observe that the interval widths are smallest for areal units that border many other units and largest for areal units with very few neighbors. The neighborhoods with smallest widths are the parks (Fairmount, Wissahickon and Pennipack) which are surrounded by many block groups due to their large surface area. The neighborhoods with largest widths are at the border of the city with only one or two neighboring units. This phenomenon is more striking for the trends in crime over time ($\beta_i$'s) for which there is a less strong signal in the data compared to the mean level of crime ($\alpha_i$'s) for each neighborhood.

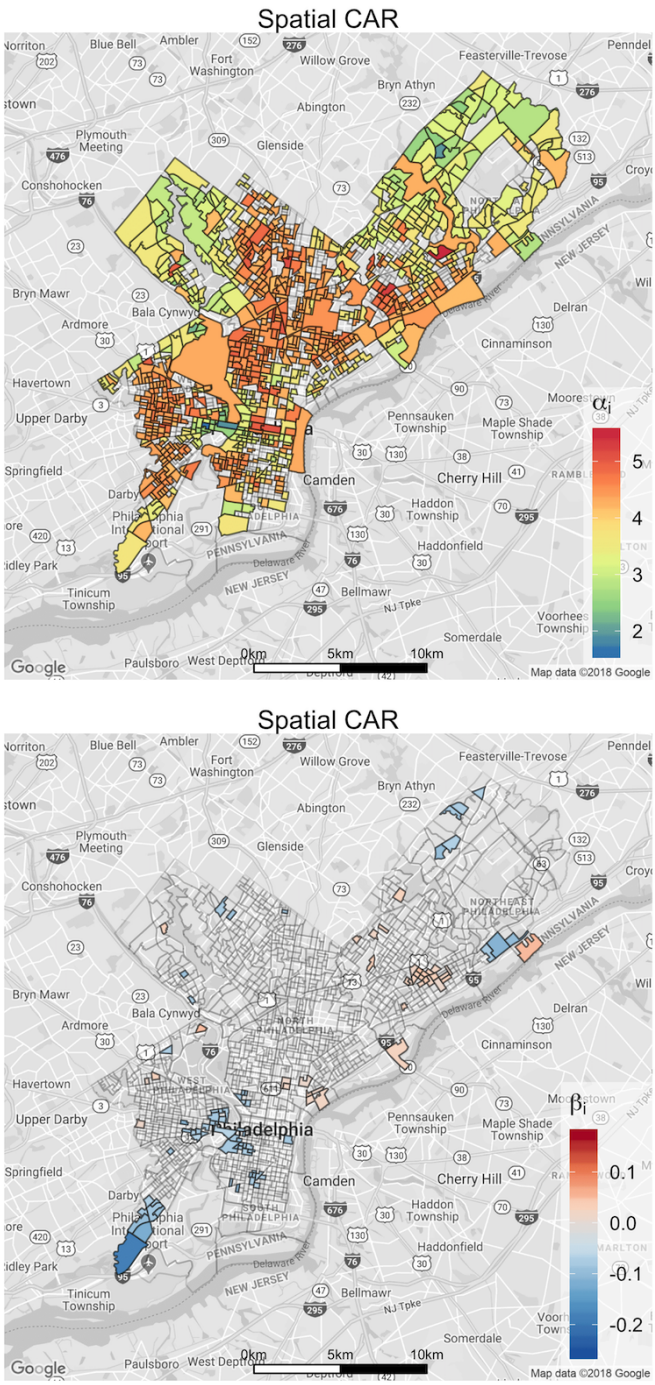


Figure 32: **Top:** The neighborhoods where the 95% credible interval for $\alpha_i$ does not contain the global mean level of crime. **Bottom:** The neighborhoods where the 95% credible interval for $\beta_i$ oes not contain the global time trend in crime. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

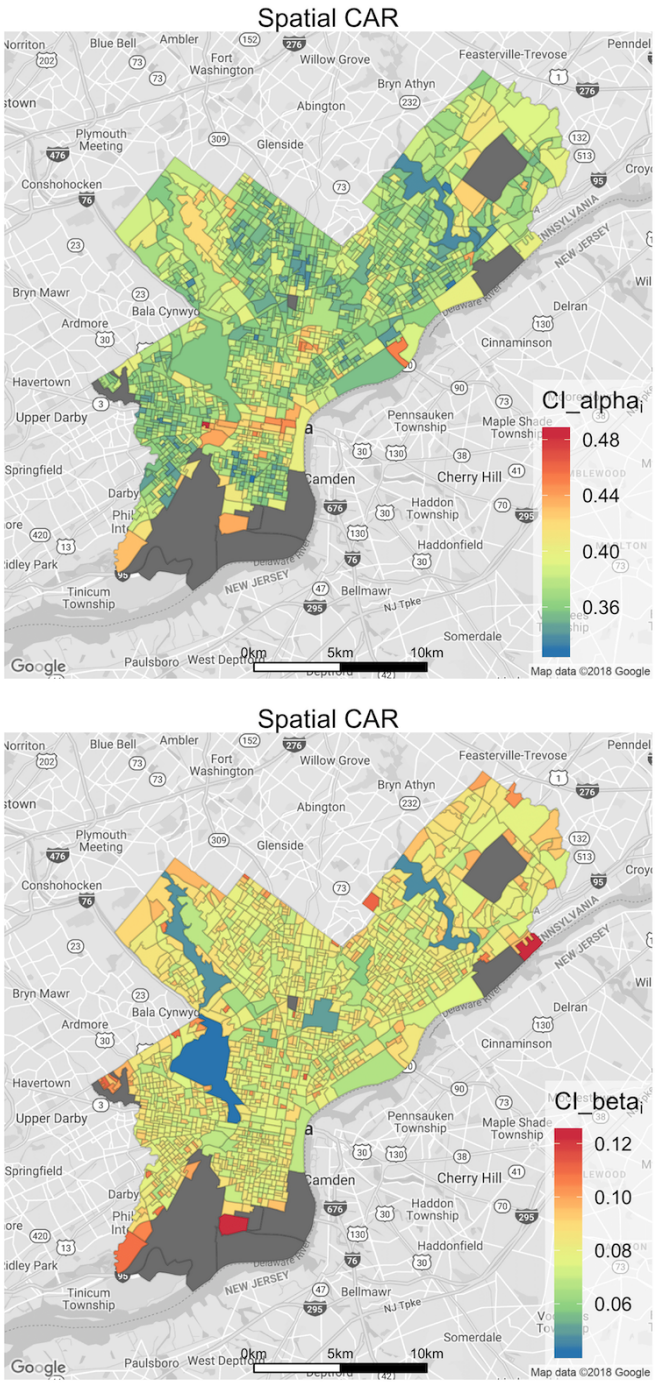


Figure 33: **Top:** The width of the 95% credible intervals for $\alpha_i$. **Bottom:** The width of the 95% credible intervals for $\beta_i$. These maps were created with the R package `ggmap` (Kahle and Wickham, 2013).

# APPENDIX B

# Bayesian Clustering with Particle Optimization

## B.1. Proof of Proposition 1

In this Section 3.1 we state that we can find the set of $L$ particles with largest posterior by finding a variational approximation of the tempered posterior $\Pi_\lambda$. Here we restate Proposition 1 and provide the proof.

Remember that we denote with $\Gamma_L = \{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\}$ the set of $L$ particles with largest posterior mass, with $q(\cdot \mid \Gamma, \mathbf{w})$ the discrete distribution that places probability $w_\ell$ on the particle $\boldsymbol{\gamma}_\ell$ and with $\mathcal{Q}_L$ the collection of all such distributions supported on at most $L$ particles. Moreover, for each $\lambda > 0$, let $\pi_\lambda$ be the mass function of the tempered marginal posterior $\Pi_\lambda$, where $\pi_\lambda(\boldsymbol{\gamma}) \propto \pi(\boldsymbol{\gamma} \mid \mathbf{y})^{\frac{1}{\lambda}}$.

**Proposition 2.** *Suppose that $\pi(\boldsymbol{\gamma} \mid \mathbf{y})$ is supported on at least $L$ distinct particles and that $\pi_\lambda(\boldsymbol{\gamma}) \neq \pi_\lambda(\boldsymbol{\gamma}')$ for $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}'$. Let $q^\star_\lambda(\cdot|\Gamma^\star(\lambda), \mathbf{w}^\star(\lambda))$ be the distribution in $\mathcal{Q}_L$ that is closest to $\Pi_\lambda$ in a Kullback-Leibler sense:*

$$q^\star_\lambda = \underset{q \in \mathcal{Q}_L}{\arg\min} \left\{ \sum_{\boldsymbol{\gamma}} q(\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\gamma})}{\pi_\lambda(\boldsymbol{\gamma})} \right\}.$$

*Then $\Gamma^\star(\lambda) = \Gamma_L$ and for each $\ell = 1, \ldots, L$, $w^\star_\ell(\lambda) \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{y})^{\frac{1}{\lambda}}$*

*Proof.* Denote the optimal particles $\Gamma^\star(\lambda) = \{\boldsymbol{\gamma}^\star_1, \ldots, \boldsymbol{\gamma}^\star_{L^\star}\}$. Straightforward calculus verifies that $w^\star_\ell(\lambda) \propto \pi_\lambda(\boldsymbol{\gamma}^\star_\ell)$. We thus compute

$$\mathrm{KL}(q^\star \parallel \pi_\lambda) = \sum_{\boldsymbol{\gamma}} q^\star(\boldsymbol{\gamma}) \log \frac{q^\star(\boldsymbol{\gamma})}{\pi_\lambda(\lambda)} = -\log \Pi_\lambda(\Gamma^\star(\ell))$$

Since $\Pi_\lambda$ is supported on at least $L$ models, we see from this computation that if $\Gamma^\star$ contained

fewer than $L$ particles, we could achieve a lower Kullback-Leibler divergence by adding another particle $\tilde{\boldsymbol{\gamma}}$ not currently in $\Gamma^\star$ that has positive $\Pi_\lambda$-probability to the particle set and updating the importance weights $\mathbf{w}$ accordingly.

Now if $\Gamma^\star$ contains $L$ models but $\Gamma^\star(\lambda) \neq \Gamma_L$, we know $\Pi_\lambda(\Gamma^\star(\lambda)) < \Pi_\lambda(\Gamma_L)$. Thus, replacing $\Gamma^\star(\lambda)$ by $\Gamma_L$ and adjusting the importances weights accordingly would also result in a lower Kullback-Liebler divergence. □

## B.2. Various hyper-parameter choices

The main model described in Section 2 depends on several hyper-parameters, which need to be fixed by the practitioner: the parameters for the prior for $\sigma$ ($\nu_\sigma$ and $\lambda_\sigma$) and the multiplicative constants to specify within and between cluster variance ($a_1, a_2, b_1$ and $b_2$). We will now describe the heuristic used to specify such values.

Let us consider each neighborhood separately and fit a simple linear regression model in each one: let $\hat{\alpha}_i$ and $\hat{\beta}_i$ be the least square estimates and $\hat{\sigma}_i^2$ be the estimated residual variance for neighborhood $i$. Since these estimates do not incorporate any prior information or sharing of information, we can think of them as an approximation of $\alpha_i, \beta_i$ given the partition with $N$ clusters $\gamma_N$; in fact under such configuration the coefficients are exchangeable and the only shrinkage induced is through the common variance parameter. Given this, one heuristic desideratum is that the marginal prior on $\boldsymbol{\alpha} \mid \gamma = \gamma_N$ should assign substantial probability to range of the $\hat{\alpha}_i$. Specifically, we will make sure that this conditional prior places 95% of its probability over the range of the $\hat{\alpha}_i$'s. Since $\boldsymbol{\alpha} \mid \gamma = \gamma_N \sim N(0, \sigma^2(a_1/(1-\rho) + a_2)I_n)$, we constrain $a_1$ and $a_2$ so that

$$\frac{a_1}{1-\rho} + a_2 = \frac{\max_i |\hat{\alpha}_i|^2}{4\hat{\sigma^2}}.$$

In order to determine each of $a_1$ and $a_2$, we need a second constraint. To this end, consider

the highly stylized setting in which we have $K$ overlapping clusters with equal variance $\sigma^2_{\text{cl}}$ whose means are equally spaced at distance $2\sigma_{cl}$. The idea of this second heuristic is to match such a stylized description to the observe distribution of $\hat{\alpha}_i$. In essence, this involves covering the range of $\hat{\alpha}_i$ with $K+1$ "chunks" of length $2\sigma_{\text{cl}}$. While the exact value of $\sigma_{\text{cl}}$ is unknown, we have found it useful to approximate it $a_1\sigma^2/(1-\rho)$. This approximation tends to produce smaller values of $a_1$, which in turn encourages a relatively small number of clusters.

With these two constraints we find:

$$a_1 = \frac{(\max(\hat{\alpha}_i) - \min(\hat{\alpha}_i))^2}{4(K+1)^2\hat{\sigma}^2/(1-\rho)}$$
$$a_2 = \frac{\max_i |\hat{\alpha}_i|^2}{4\hat{\sigma^2}} - \frac{a_1}{1-\rho}.$$

Similarly for the $\hat{\beta}_i$'s we find:

$$b_1 = \frac{(\max(\hat{\beta}_i) - \min(\hat{\beta}_i))^2}{4(K+1)^2\hat{\sigma}^2/(1-\rho)}$$
$$b_2 = \frac{\max_i |\hat{\beta}_i|^2}{4\hat{\sigma^2}} - \frac{b_1}{1-\rho}.$$

In order to operationalize these heuristics, we must specify an initial guess at $K$. We have found in our experiments, setting $K = \lfloor \log N \rfloor$ works quite well. It, moreover, accords with the general behavior of the Ewens-Pitman prior.

Finally, to specify the prior for $\sigma^2$ we can use the collection of $\hat{\sigma}^2_i$'s: by matching mean and variance, we can recover $\nu_\sigma = 2\frac{m^2}{v} + 4$ and $\lambda_\sigma = m(1 - \frac{2}{\nu_\sigma})$, where $m$ and $v$ are the empirical mean and variance of the $\hat{\sigma}^2_i$'s.

## B.3. Additional Synthetic Data Evaluation

In Section 4, we generated several synthetic datasets based on a 20 grid of census tracts partitioned into four clusters of size 12, 188, 100, and 100, as seen in Figure 9. Within each cluster, we drew the $\alpha_i$'s from a CAR model centered at a specified cluster mean with $\rho = 0.95$ and variance scale 0.2. Across the different specifications of cluster means, we always fixed the cluster mean of the 12-tract "cross" and the 100 tract square in the upper right corner to be zero. We then fixed the mean of the 188-tract cluster on the left hand side to be $-\Delta$ and the mean of the 100-tract cluster in the lower right corner to be $\Delta$. We generated datasets for each of $\Delta = 0, 1, \ldots, 5$. The high, medium, and low separation settings in Figure 3 and 4 correspond to $\Delta = 5, 3$, and 1, respectively.

In Section 3.4, we compared the partition selection performance of our method to that of k-means and spectral clustering. Figure 34 shows the estimated partitions from k-means and spectral clustering on the same dataset used to generate Figure 4. Across these datasets, the optimal number of clusters for k-means was always three, according to the "elbow method." However, because k-means does not implicitly account for our spatial connectedness constraints, we post-processed the recovered partition by treating disconnected parts of clusters identified by k-means as their own separate clusters.

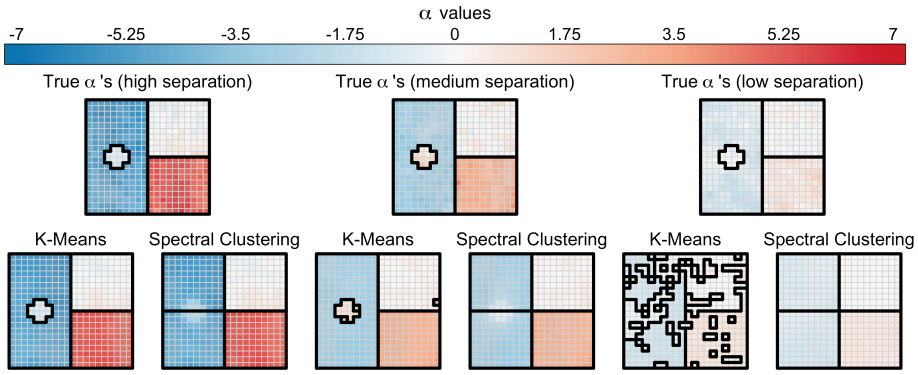


Figure 34: Partitions recovered by k-means and spectral clustering for three different cluster separation settings. The color of each tract corresponds to the estimated parameter value $\mathbb{E}[\alpha_i \mid \mathbf{y}, \boldsymbol{\gamma}]$.

## B.4. Additional Results for Clustering in Philadelphia

In figure 35 we represent the best three particles recovered by the models where the priors are specified as Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ and Uniform on $\mathcal{SP}$ for $\gamma^{(\beta)}$ (top panel) and Uniform prior on $\mathcal{SP}$ for $\gamma^{(\beta)}$ and Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ (bottom panel).

## B.5. Derivation of Closed Form Expressions

### B.5.1. One Partition Derivations

In Section 4, we considered a simpler model, in which we ignored the time trend and only focused on clustering the intercepts. That model was:

$$\begin{aligned}
\boldsymbol{\gamma} = \{S_1, \ldots, S_K\} &\sim \mathcal{P}_{\boldsymbol{\gamma}} \\
\sigma^2 &\sim \text{Inv. Gamma}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
\overline{\alpha}_k | \sigma^2 &\sim N(0, a_2 \sigma^2) \;\; \text{for each } k = 1, \ldots, K \\
\boldsymbol{\alpha}_{S_k} | \overline{\alpha}_k, \sigma^2 &\sim N_{n_k}(\overline{\alpha} \mathbf{1}_{n_k}, a_1 \sigma^2 \Sigma_k^{(\alpha)}) \;\; \text{for each } k = 1, \ldots, K \\
y_{i,t} | \alpha_i, \sigma^2 &\sim N(\alpha_i, \sigma^2) \;\; \text{for each } i = 1, \ldots, N, \text{ and } t = 1, \ldots, T
\end{aligned}$$

For the sake of completeness, we derive the corresponding marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ and conditional expectation $\mathbb{E}[\boldsymbol{\alpha} \mid \boldsymbol{\gamma}, \mathbf{y}]$ for this simpler setting.

Now observe

$$\begin{aligned}
p(\mathbf{y} | \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\gamma}) &\propto \prod_{k=1}^{K} \prod_{i \in S_k} \left(\sigma^2\right)^{-\frac{T}{2}} \exp\left\{ -\frac{T(\overline{y}_i - \alpha_i)^2 + (T-1)s_i^2}{2\sigma^2} \right\} \\
&\propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{ -\frac{(T-1)\sum_{i=1}^{N} s_i^2}{2\sigma^2} \right\} \prod_{k=1}^{K} p(\overline{\mathbf{y}}_{S_k} | \boldsymbol{\alpha}_{S_k}, \sigma^2, \boldsymbol{\gamma})
\end{aligned}$$

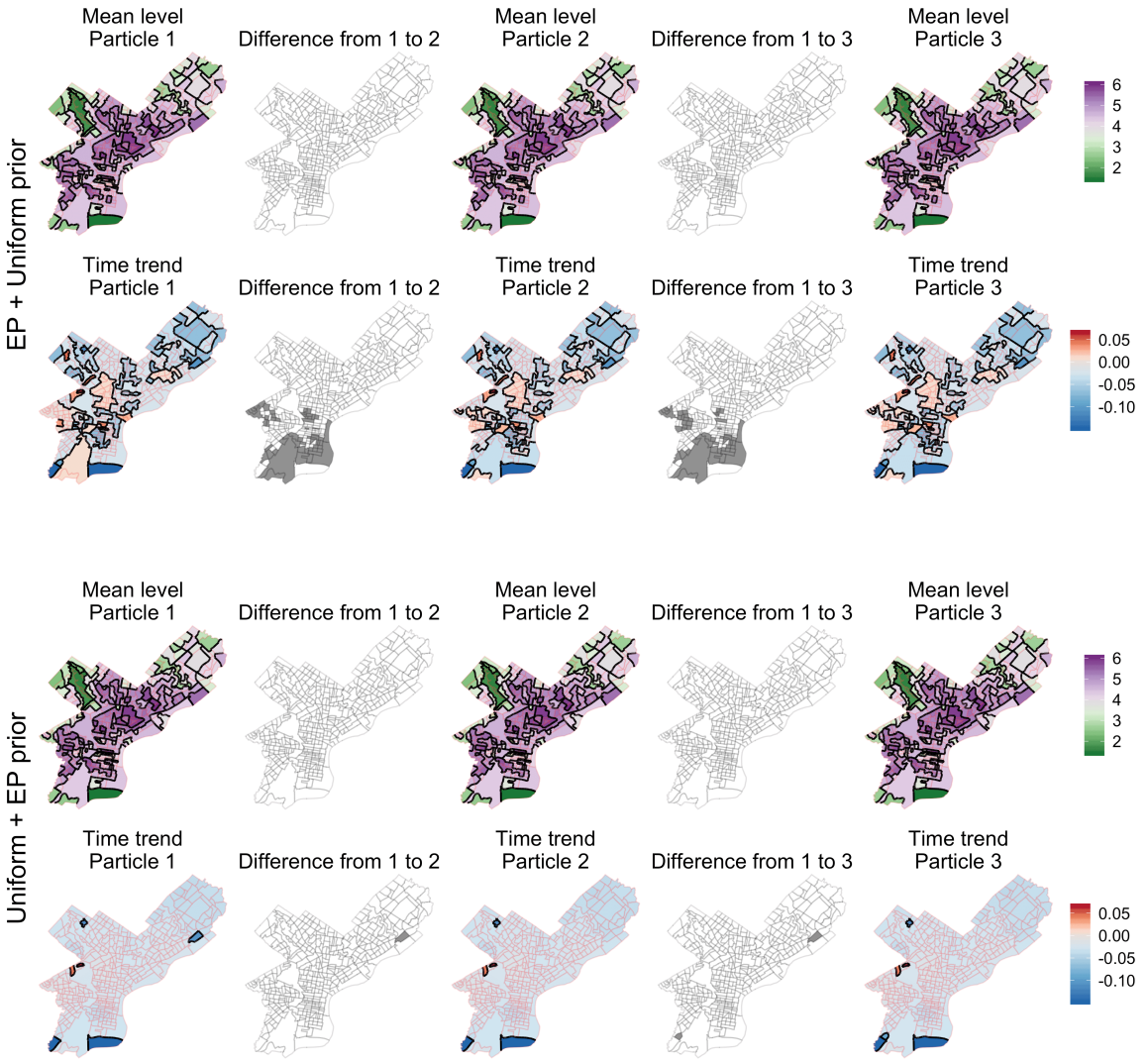


Figure 35: Colored plots: Top three models identified by our procedure. The thick borders represent the partition, and the color represents the posterior mean of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Black and white plots: transition from the model on the left to the model on the right. The greyed areas represent the neighborhoods whose cluster assignments change in the partitions on the sides. **Top:** Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$ and Uniform on $\mathcal{SP}$ for $\gamma^{(\beta)}$. **Bottom:** Uniform prior on $\mathcal{SP}$ for $\gamma^{(\beta)}$ and Ewens-Pitman prior with $\eta = 5$ for $\gamma^{(\alpha)}$.

where $\overline{\mathbf{y}}_{S_k}|\boldsymbol{\alpha}_{S_k}, \sigma^2, \boldsymbol{\gamma} \sim N_{n_k}(\boldsymbol{\alpha}_{S_k}, T^{-1}\sigma^2 I_{n_k})$. From here, we conclude

$$p(\overline{y}|\sigma^2, \boldsymbol{\gamma}) \propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{-\frac{(T-1)\sum_{i=1}^N s_i^2}{2\sigma^2}\right\} \prod_{k=1}^K p(\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma})$$

To derive $p(\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma})$, we first note that marginally

$$\boldsymbol{\alpha}_{S_k}|\sigma^2 \sim N_{n_k}(0 \cdot \mathbf{1}_{n_k}, \sigma^2[a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top]).$$

Now marginalizing out $\boldsymbol{\alpha}_{S_k}$ we have

$$\overline{\mathbf{y}}_{S_k}|\sigma^2, \boldsymbol{\gamma} \sim N_{n_k}\left(0\mathbf{1}_{n_k}, \sigma^2\left[a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top + T^{-1}I_{n_k}\right]\right)$$

Hence

$$\begin{aligned} p(\mathbf{y}|\sigma^2, \boldsymbol{\gamma}) \propto \left(\sigma^2\right)^{-\frac{N(T-1)}{2}} \exp\left\{-\frac{(T-1)\sum_{i=1}^N s_i^2}{2\sigma^2}\right\} \\ \times \prod_{k=1}^K \left(\sigma^2\right)^{-\frac{n_k}{2}} |\Omega_k^{(y)}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^K \overline{\mathbf{y}}_k^\top \Omega_k^{(y)} \overline{\mathbf{y}}_k\right\} \end{aligned}$$

where $\Omega_k^{(y)} = [a_1\Sigma_k^{(\alpha)} + a_2\mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top + T^{-1}I_{n_k}]^{-1}$.

Marginalizing out $\sigma^2$, we conclude

$$p(\overline{\mathbf{y}}|\boldsymbol{\gamma}) = C(N, \nu_\sigma, \lambda_\sigma) \times \left(\prod_{k=1}^K |\Omega_k^{(y)}|\right)^{\frac{1}{2}} \times \left[\frac{\nu_\sigma\lambda_\sigma}{2} + \frac{1}{2}\sum_{k=1}^K \overline{\mathbf{y}}_k^\top \Omega_k^{(\alpha)} \overline{\mathbf{y}}_k + \frac{(T-1)}{2}\sum_{i=1}^N s_i^2\right]^{-\frac{\nu_\sigma + NT}{2}}$$

We further compute

$$p(\overline{y}_{S_k}, \boldsymbol{\alpha}_{S_k}|\sigma^2, \boldsymbol{\gamma}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\alpha}_{S_k}^\top V^{-1}\boldsymbol{\alpha}_{S_k} - 2\boldsymbol{\alpha}_{S_k}^\top T\overline{\mathbf{y}}_{S_k}\right]\right\},$$

where $V^{-1} = \left[ TI_{n_k} + \left( a_1 \Sigma_k^{(\alpha)} + a_2 \mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top \right)^{-1} \right]$. From here, we immediate conclude that

$$\mathbb{E}[\boldsymbol{\alpha}_{S_k} | \overline{\mathbf{y}}_{S_k}, \boldsymbol{\gamma}] = T \times V \overline{\mathbf{y}}_{S_k}.$$

Finally, note that

$$\begin{aligned} p(\overline{\alpha}_k, \boldsymbol{\alpha}_{S_k}, \overline{\mathbf{y}}_{S_k} | \sigma^2, \boldsymbol{\gamma}) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left[ (\overline{\mathbf{y}} - \boldsymbol{\alpha}_{S_k})^\top T (\overline{\mathbf{y}} - \boldsymbol{\alpha}_{S_k})^\top \right] \right\} \\ &\times \exp\left\{ -\frac{1}{2\sigma^2} \left[ (\boldsymbol{\alpha}_{S_k} - \overline{\alpha}_k \mathbf{1}_{n_k})^\top a_1^{-1} \Omega_k^{(\alpha)} (\boldsymbol{\alpha}_{S_k} - \overline{\alpha}_k \mathbf{1}_{n_k}) \right] \right\} \\ &\times \exp\left\{ -\frac{1}{2\sigma^2} \overline{\alpha}_k^2 a_2^{-1} \right\} \end{aligned}$$

Therefore,

$$p(\overline{\alpha}_k | \boldsymbol{\alpha}_{S_k}, \overline{\mathbf{y}}, \sigma^2, \boldsymbol{\gamma}) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left[ \overline{\alpha}_k^2 \left( a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \right) - 2\overline{\alpha}_k a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \boldsymbol{\alpha}_{S_k} \right] \right\}$$

By the Woodbury identity, we compute

$$\begin{aligned} \left[ a_1 \Sigma_k^{(\alpha)} + a_2 \mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top \right]^{-1} &= a_1^{-1} \Omega_k^{(\alpha)} - a_1^{-1} \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \left[ a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k} \right]^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} a_1^{-1} \\ &= a_1^{-1} \Omega_k^{(\alpha)} - a_1^{-2} (1-\rho)^2 \times \left[ a_2^{-1} + a_1^{-1} (1-\rho) n_k \right]^{-1} \times \mathbf{1}_{n_k} \mathbf{1}_{n_k}^\top \end{aligned}$$

So the posterior conditional mean of $\overline{\alpha}_k$ is given by

$$\mathbb{E}[\overline{\alpha}_k | \boldsymbol{\alpha}_{S_k}, \mathbf{y}_{S_k}, \boldsymbol{\gamma}] = \frac{a_1^{-1} \mathbf{1}^\top \Omega_k^{(\alpha)} \boldsymbol{\alpha}_{S_k}}{a_2^{-1} + a_1^{-1} \mathbf{1}_{n_k}^\top \Omega_k^{(\alpha)} \mathbf{1}_{n_k}^\top} = \frac{a_1^{-1} (1-\rho) \mathbf{1}_{n_k}^\top \boldsymbol{\alpha}_{S_k}}{a_2^{-1} + a_1^{-1} n_k (1-\rho)}$$

*Note: observe that as $a_2 \to \infty$ (i.e. as we allow the variability of the cluster means to increase), this conditional expectation converges to the $n_k^{-1} \mathbf{1}^\top \boldsymbol{\alpha}_{S_k}$, the arithmetic mean of the parameters within each block-group.*

### B.5.2. Two Partition Derivations

Recall from Section 3.2 that our full mode is:

$$
\begin{aligned}
\gamma^{(\alpha)}, \gamma^{(\beta)} &\sim \mathrm{EP}(\eta; \mathcal{SP}) \\
\sigma^2 &\sim \mathrm{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right) \\
(\overline{\alpha}_k)_k &\overset{iid}{\sim} N(0, a_2\sigma^2) \\
(\overline{\beta}_{k'})_{k'} &\overset{iid}{\sim} N(0, b_2\sigma^2) \\
(\boldsymbol{\alpha}_k)_k &\overset{ind}{\sim} \mathrm{CAR}(\overline{\alpha}_k, a_1\sigma^2, W_k^{(\alpha)}) \\
(\boldsymbol{\beta}_{k'})_{k'} &\overset{ind}{\sim} \mathrm{CAR}(\overline{\beta}_{k'}, b_1\sigma^2, W_{k'}^{(\beta)}) \\
(y_{i,t})_{i,t} &\overset{ind}{\sim} \mathrm{N}(\alpha_i + \beta_i(t - \bar{t}), \sigma^2)
\end{aligned}
$$

We exploit the conditional conjugacy present in this model in several places. First, we have closed form expressions for the conditional posterior means $\mathbb{E}[\boldsymbol{\alpha} \mid \mathbf{y}, \boldsymbol{\gamma}]$ and $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\gamma}]$, which we use in our particle optimization procedure to propose new transitions. Second, we can compute the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\gamma})$ in closed form, which we use to evaluate the optimization objective and pick between multiple transitions. Below, we carefully derive these closed form expressions, noting that in several places, we can avoid potentially expensive matrix inversions. In particular, the choice to center the time variable, thereby ensuring an orthogonal design matrix within each neighborhood, facilitates rapid likelihood evaluations.

**Distribution of $\boldsymbol{\alpha}_k$** Let us first consider the vector of parameters $\boldsymbol{\alpha}_k$ in cluster $S_k^{(\alpha)}$ given $\sigma^2$: by marginalizing the distribution of the grand cluster mean $\overline{\alpha}_k$, we find that its distribution is a multivariate normal with covariance matrix $\sigma^2\Sigma_k^{(\alpha)}$, where $\Sigma_k^{(\alpha)} = a_1\Sigma_{k,\mathrm{CAR}}^{(\alpha)} + a_2\mathbf{1}\mathbf{1}^\top = a_1\left[\rho(W_k^{(\alpha)})^* + (1-\rho)\mathbf{I}\right]^{-1} + a_2\mathbf{1}\mathbf{1}^\top$. Note that its precision matrix

can be computed using Woodbury's formula without having to invert any matrix:

$$(\Sigma_k^{(\alpha)})^{-1} = a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} - a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)}\mathbf{1}\left(a_1^{-1}\mathbf{1}^\top\Omega_{k,\text{CAR}}^{(\alpha)}\mathbf{1} + a_2^{-1}\right)^{-1}\mathbf{1}^\top a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} =$$
$$= a_1^{-1}\Omega_{k,\text{CAR}}^{(\alpha)} - \frac{a_1^{-2}(1-\rho)^2}{a_1^{-1}n_k(1-\rho)+a_2^{-1}}\mathbf{1}\mathbf{1}^\top$$

where $\Omega_{k,\text{CAR}}^{(\alpha)} = \left(\Sigma_{k,\text{CAR}}^{(\alpha)}\right)^{-1} = \rho(W_k^{(\alpha)})^* + (1-\rho)\mathbf{I}$; the second line follows from noticing that $\mathbf{1}$ is both a left and right eigenvector of $\Omega_{k,\text{CAR}}^{(\alpha)}$ with eigenvalue $1-\rho$. Similarly this holds for the distribution of $\boldsymbol{\beta}_{k'}$.

**Distribution of $\boldsymbol{\alpha}$** Next, we can write the distribution of the whole vector $\boldsymbol{\alpha}$ given $\sigma^2$ and $\gamma^{(\alpha)}$: by combining the distributions of the cluster specific parameters $\boldsymbol{\alpha}_k$'s, and using the independence between different clusters, we find that the distrubution of $\boldsymbol{\alpha}$ given $\sigma^2$ and $\gamma^{(\alpha)}$ is a multivariate normal with mean zero and covariance matrix that can be found by combining the $\Sigma_k^{(\alpha)}$'s. Because of the independence between clusters, *there exists an ordering of the indices of* $\boldsymbol{\alpha}$ so that the covariance matrix of $\boldsymbol{\alpha}|\gamma_\alpha, \sigma^2$ has a block-diagonal structure. We denote such permutation of the indices with $\pi^{(\alpha)}$, and it can be constructed by mapping the first $n_1$ elements to the indices in the first cluster ($\{\pi^{(\alpha)}(1), \ldots, \pi^{(\alpha)}(n_1)\} = S_1^{(\alpha)}$), the following $n_2$ elements to the indices in the second cluster ($\{\pi^{(\alpha)}(n_1+1), \ldots, \pi^{(\alpha)}(n_1+n_2)\} = S_2^{(\alpha)}$), and so on. With such ordering, the $k$th diagonal block of the covariance matrix is $\sigma^2\Sigma_k^{(\alpha)}$. Similarly, we can find a (potentially different) permutation $\pi^{(\beta)}$ for $\boldsymbol{\beta}$ and derive the distribution of $\boldsymbol{\beta}_\pi|\sigma^2, \gamma^{(\beta)}$.

**Notation** To describe the distributions of interest we can represent our model in the form of a unique linear model, by combining all the observations in a vector $Y$, combining the reodered coefficients in a unique vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}_\pi, \boldsymbol{\beta}_\pi)$ and appropriately constructing the covariate matrix $X$. In the next paragraphs we will provide with the details on how we constructed such vectors and matrix.

To build the column vector $Y$ we stack the vectors $\mathbf{y}_i$ with $i = 1, \ldots, N$: $Y$ is a vector

of length $N \cdot T$ and each block of $T$ rows corresponds to a particular neighborhood; in particular, the $((i-1)T+t)$th entry of $Y$ corresponds to $y_{i,t}$.

The vector of coefficients $\boldsymbol{\theta}$ is found by concatenating the reordered $\boldsymbol{\alpha}_\pi$ and $\boldsymbol{\beta}_\pi$: for $i = 1, \ldots, N$, elements $\theta_i = \alpha_{\pi^{(\alpha)}(i)}$ and $\theta_{N+i} = \beta_{\pi^{(\beta)}(i)}$.

The matrix of covariates $X$ then has dimensions $NT \times 2N$; each block of $T$ rows corresponds to a neighborhood and each column corresponds to an element of $\boldsymbol{\theta}$: the first $N$ columns correspond to the elements of $\boldsymbol{\alpha}_\pi$ and the second $N$ columns to $\boldsymbol{\beta}_\pi$. The rows of $X$ corresponding to neighborhood $i$ (rows $(i-1)T+t$ with $t = 1, \ldots T$) have an element equal to 1 in the $(\pi^{(\alpha)})^{-1}(i)$th column, an element equal to $x_{it} = t - \bar{t}$ in the $(N + (\pi^{(\beta)})^{-1}(i))$th column, and zero elsewhere. With such construction, the $(i-1)T+t$ row of the equation $Y = X\boldsymbol{\theta}$ corresponds to $y_{i,t} = \theta_{(\pi^{(\alpha)})^{-1}(i)} + x_{it}\theta_{N+(\pi^{(\beta)})^{-1}(i)} = \alpha_i + (t-\bar{t})\beta_i$.

**Marginal likelihood** $Y|\gamma^{(\alpha)}, \gamma^{(\beta)}$ To recover the marginal likelihood $p(Y|\gamma^{(\alpha)}, \gamma^{(\beta)})$ we compute

$$\begin{aligned}
&\int \left[ \int p(Y|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\alpha}|\gamma^{(\alpha)}, \sigma^2) p(\boldsymbol{\beta}|\gamma^{(\beta)}, \sigma^2) d\boldsymbol{\alpha} d\boldsymbol{\beta} \right] p(\sigma^2) d\sigma^2 = \\
&= \int \left[ \int p(Y|\boldsymbol{\alpha}_\pi, \boldsymbol{\beta}_\pi, \sigma^2) p(\boldsymbol{\alpha}_\pi|\gamma^{(\alpha)}, \sigma^2) p(\boldsymbol{\beta}_\pi|\gamma^{(\beta)}, \sigma^2) d\boldsymbol{\alpha}_\pi d\boldsymbol{\beta}_\pi \right] p(\sigma^2) d\sigma^2 = \\
&= \int \left[ \int p(Y|\boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) d\boldsymbol{\theta} \right] p(\sigma^2) d\sigma^2.
\end{aligned}$$

Let us first compute $p(Y|\sigma^2, \gamma^{(\alpha)}, \gamma^{(\beta)}) = \int p(Y|\boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) d\boldsymbol{\theta}$. Using the notation for linear regression we can write $p(Y|\boldsymbol{\theta}, \sigma^2) = N(X\boldsymbol{\theta}, \sigma^2\mathbf{I})$. The prior for $\boldsymbol{\theta}$ is a normal distribution with mean zero and block covariance matrix $\Sigma_\theta$: the first $n \times n$ block corresponds to the covariance matrix of $\boldsymbol{\alpha}$ and the second to the one for $\boldsymbol{\beta}$.

By integrating out $\boldsymbol{\theta}$, $p(Y|\gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^2) = N\left(\mathbf{0}, \sigma^2\Sigma_Y\right)$ where $\Sigma_Y = \mathbf{I} + X\Sigma_\theta X^\top$. Its precision matrix can be computed using Woodbury's formula again: $\Sigma_Y^{-1} = \mathbf{I} - X(\Sigma_\theta^{-1} + X^\top X)^{-1}X^\top$. Note that $X^\top X$ is a diagonal matrix, and we derive its form at the end of this chapter.

The marginal likelihood can now be derived by integrating out $\sigma^2$:

$$
\begin{aligned}
p(Y|\gamma^{(\alpha)},\gamma^{(\beta)}) &= \int p(Y|\sigma^2,\gamma^{(\alpha)},\gamma^{(\beta)})p(\sigma^2)d\sigma^2 = \\
&= \pi^{-nT/2}\det(\Sigma_Y)^{-1/2}\frac{(\nu_\sigma\lambda_\sigma/2)^{\nu_\sigma/2}}{\Gamma(\frac{\nu_\sigma}{2})}\int(\sigma^2)^{-\frac{NT+\nu_\sigma}{2}-1}e^{-\frac{Y^\top\Sigma_Y^{-1}Y+\nu_\sigma\lambda_\sigma}{2\sigma^2}}d\sigma^2 = \\
&= \pi^{-nT/2}\det(\Sigma_Y)^{-1/2}\frac{\Gamma(\frac{NT+\nu_\sigma}{2})}{\Gamma(\frac{\nu_\sigma}{2})}\left(\frac{\nu_\sigma\lambda_\sigma}{2}\right)^{\nu_\sigma/2}\left(\frac{\nu_\sigma\lambda_\sigma+Y^\top\Sigma_Y^{-1}Y}{2}\right)^{-(NT+\nu_\sigma)/2} = \\
&= \pi^{-nT/2}\det(\Sigma_Y)^{-1/2}\frac{\Gamma(\frac{NT+\nu_\sigma}{2})}{\Gamma(\frac{\nu_\sigma}{2})}\left(\frac{\nu_\sigma\lambda_\sigma}{2}\right)^{-NT/2}\left(1+\frac{Y^\top\Sigma_Y^{-1}Y}{\nu_\sigma\lambda_\sigma}\right)^{-(NT+\nu_\sigma)/2}.
\end{aligned}
$$

Note that if $\lambda_\sigma = 1$, this is multivariate t-distribution with $\nu_\sigma$ degrees of freedom.

For this we need to compute the quadratic form

$$Y^\top\Sigma_Y^{-1}Y = Y^\top Y - Y^\top X(\Sigma_\theta^{-1}+X^\top X)^{-1}X^\top Y.$$

Because of the block diagonal structure of $\Sigma_\theta^{-1}+X^\top X$ we can write this as a sum over the clusters of the two partitions. Consider the column vector $X^\top Y$ of length $2N$: the first $N$ elements correspond to the summary statistics related to the $\alpha_{\pi(i)}$'s and we will denote the ones corresponding to cluster $S_k^{(\alpha)}$ with $(X^\top Y)_k^{(\alpha)}$, while the second $N$ elements are for the $\beta_i$'s and we denote with $(X^\top Y)_{k'}^{(\beta)}$ the ones for cluster $S_{k'}^{(\beta)}$ . Now we can write

$$
\begin{aligned}
Y^\top X(\Sigma_\theta^{-1}+X^\top X)^{-1}X^\top Y = &\sum_{k=1}^{K^{(\alpha)}}(X^\top Y)_k^{(\alpha)\top}((\Sigma_k^{(\alpha)})^{-1}+T\mathbf{I})^{-1}(X^\top Y)_k^{(\alpha)} \\
&+\sum_{k'=1}^{K^{(\beta)}}(X^\top Y)_{k'}^{(\beta)\top}((\Sigma_{k'}^{(\beta)})^{-1}+\textstyle\sum x_t^2\mathbf{I})^{-1}(X^\top Y)_{k'}^{(\beta)}
\end{aligned}
$$

where $(\Sigma_k^{(\alpha)})^{-1}+T\mathbf{I}$ is the diagonal blocks of $\Sigma_\theta^{-1}+X^\top X$ corresponding to cluster $S_k^{(\alpha)}$ and $(\Sigma_{k'}^{(\beta)})^{-1}+\sum x_t^2\mathbf{I}$ corresponds to $S_{k'}^{(\beta)}$; each of them can be inverted using methods for symmetric positive definite matrices.

To compute the marginal likelihood we are left we calculating the determinant of $\Sigma_Y$, where

we can use the reciprocal of the determinant of its inverse

$$\det(\Sigma_Y^{-1}) = \det(\mathbf{I} - X(\Sigma_\theta^{-1} + X^\top X)^{-1}X^\top) = \det(\mathbf{I} - (\Sigma_\theta^{-1} + X^\top X)^{-1}X^\top X)$$

where the last equality is given by Sylvester's formula, and allows us to compute the determinant of a smaller dimensional matrix. Moreover, because of its block diagonal structure, we can compute the determinant block-wise.

**Posterior mean of $\boldsymbol{\alpha}, \boldsymbol{\beta}$** The calculations for the posterior mean of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are very similar: using the same notation and the results for linear regression, we can find

$$\mathbb{E}\left[\boldsymbol{\theta}|Y, \gamma^{(\alpha)}, \gamma^{(\beta)}, \sigma^{-1}\right] = \left(X^\top X + \Sigma_\theta^{-1}\right)^{-1} X^\top Y$$

and since this does not depend on $\sigma^2$, it coincides with $\mathbb{E}\left[\boldsymbol{\theta}|Y, \gamma^{(\alpha)}, \gamma^{(\beta)}\right]$. Because of the block diagonal structure of the matrices involved, we can compute the estimate of the parameter for each cluster independently. Moreover, note that the inverse of $X^\top X + \Sigma_\theta^{-1}$ is computed in the likelihood calculation, so it can be stored and does not need to be computed two times.

**Derivation of $X^\top X$** Since in our formulation the covariates are orthogonal, i.e. $\sum_{t=1}^{T} x_{it} = 0$ for all $i$, $X^\top X$ is a diagonal matrix. Note that column $X_{(\pi^{(\alpha)})^{-1}(i')}$ contains $T$ 1's in rows $t + (i'-1)\times$T and zeros elsewhere; similarly column $X_{N+(\pi^{(\beta)})^{-1}(i')}$ contains elements $(x_{i't})$ in rows $t + (i'-1) \times T$ and zero's elsewhere. Thus, when we compute $(X^\top X)_{ij}$ we consider the cross product of columns $X_i$ and $X_j$. Depending on the value of $i$ and $j$, we have the following cases:

- if $i = j \leq N$, then $(X^\top X)_{ij} = T$,
- if $i = j \geq N$, then $(X^\top X)_{ij} = \sum_t x^2_{\pi^{(\beta)}(j-N),t}$,
- if $i \leq N$ and $j = N + i$, then $(X^\top X)_{ij} = \sum_t x_{\pi^{(\beta)}(i),t} = 0$,

- if $j \leq N$ and $i = N + j$, then $(X^\top X)_{ij} = \sum_t x_{\pi^{(\beta)}(j),t} = 0$,
- for any other $i, j$, $(X^\top X)_{ij} = 0$.

Thus the matrix $X^\top X$ is a diagonal matrix: the first $n \times n$ diagonal block is $T\mathbf{I}$, and the second diagonal block is a diagonal matrix whose entries are $\sum_{t=1}^T x_{it}^2$; when we have fixed design, $x_{it} = x_t = t - \bar{t}$, then $\sum_{t=1}^T x_{it}^2 = \sum_{t=1}^T (t - \bar{t})^2$ is constant, so the second diagonal block is $\sum x_{it}^2 \mathbf{I}$. Because of the orthogonality of the covariates, the upper-right and lower-left blocks are zero matrices, since $\sum_{t=1}^T x_{it} = 0$.

**Note on cluster-wise update of calculations.** In our greedy search when we perform a move only one or two clusters in only one partition is changed: in a *split* move for $\gamma^{(\cdot)}$, a cluster is divided into two sub-clusters, and the original cluster replaced by the first, while the second creates an additional cluster; in a *merge* move, one of two clusters is deleted and the other is replaced to the merge of the two original clusters. In each case, we need to update the value of the marginal likelihood, of the prior for $\gamma^{(\cdot)}$ and of the estimate of the parameters.

Because of the block structure given by orthogonality of covariates and by the reordering of the parameters, changing the structure of some clusters does not affect the parameter estimates for other clusters that are not involved in the move. This implies that updates for updates to $S_k^{(\alpha)}$ do not affect the parameter estimates $\boldsymbol{\alpha}_h$ for $h \neq k$ or $\boldsymbol{\beta}_{k'}$ for any $k'$. Similarly, since the quadratic form $Y^\top \Sigma_Y^{-1} Y$ can be written as sum of cluster-specific quadratic forms, we can update only the quadratic form of the clusters affected and we can compute the determinant of the blocks of $\Sigma_Y$ corresponding to the modified clusters.

This allows us to invert matrices that scale like the size of the clusters, reducing the computational costs dramatically.

# APPENDIX C

# Prior Choice for Clustering Areal Data

## C.1. Hyper-parameter choice

Both the Areal-PPMx and the PPMx-SBM require the choice of the hyper-parameters for the prior distribution of the auxiliary model. In this section we describe the heuristics to choose them.

**Stochastic block model** When the auxiliary model is the Stochastic Block Model, we assume that within-cluster and between-cluster edges are i.i.d. bernoulli distributed respectively with probability $\xi_1$ and $\xi_0$. The prior for $\xi_1$ is Beta$(a, b)$ and for $\xi_0$ is Beta$(c, d)$. While we don't want these priors to be excessively informative, we would like to include in this model some prior information: firstly, we want to include the idea that $xi_1 > \xi_0$; secondly, assuming that the cohesion function is coupled with the DP distribution, we can use some properties of the DP distribution such as the average number of clusters to estimate the prior proportion of within-cluster and between-cluster connections.

To specify the prior for $\xi_1$ we can consider the prior number of successes (within-cluster connections) and of failures (within-cluster non-connections). Under the DP distribution the (asymptotic) average number of clusters is $\alpha \log(n)$, thus we can approximate the average size of a cluster with $s(n) = \frac{n}{\alpha \log(n)}$. Moreover we can consider the average degree, i.e. the average number of edges for each node, $d(n) = \frac{\sum_{ij} w_{ij}}{n}$. We can express the prior number of successes as the average cluster size times the average degree, divided by 2, since we don't want to count edges twice

$$a = s(n)d(n)/2.$$

As for the number of failures, i.e. the number of non-connections within a cluster, we can consider that, if a unit has $d(n)$ connections with other units in the same cluster, then it

has $s(n) - 1 - d(n)$ non-connections with units in the same cluster, on average. Thus we can estimate

$$b = s(n)(s(n) - 1 - d(n))/2.$$

The choice for the prior of $\xi_0$ is instead harder. In fact, it requires knowledge of the borders between clusters, which cannot be easily deduced. With a strong assumption we consider a square grid, with squared clusters. Then the border between two squares is the length of its side, which can be computed as $\sqrt{(n/\log(n))}$. Assuming that units on the border only have one connection to some other unit in another cluster we can estimate

$$c = \sqrt{n/\log(n)}$$

As for the prior number of failures (the number of between-cluster non-connections) we can use an over-estimate by just considering all the possible pair of units:

$$d = n(n-1)/2.$$

**Note on coherence** If the hyper-parameters for the auxiliary model depend on $n$, the property of coherence over sample sizes is no longer satisfied or meaningful. In the heuristic described we considered the case in which the number of units is fixed.

**Areal-PPMx** When we use the Beta-Binomial to model the number of within-cluster connections vs the total number of connections, we need to specify the prior for the probability that an edge is within-cluster (rather than between-cluster). The probability $\xi$ is modeled as Beta$(a, b)$, and we need to choose the hyper-parameters. The prior number of successes $a$ can be specified as the average cluster size times the average degree, divided by 2, as for $a$ in the SBM auxiliary model.

$$a = s(n)d(n)/2.$$

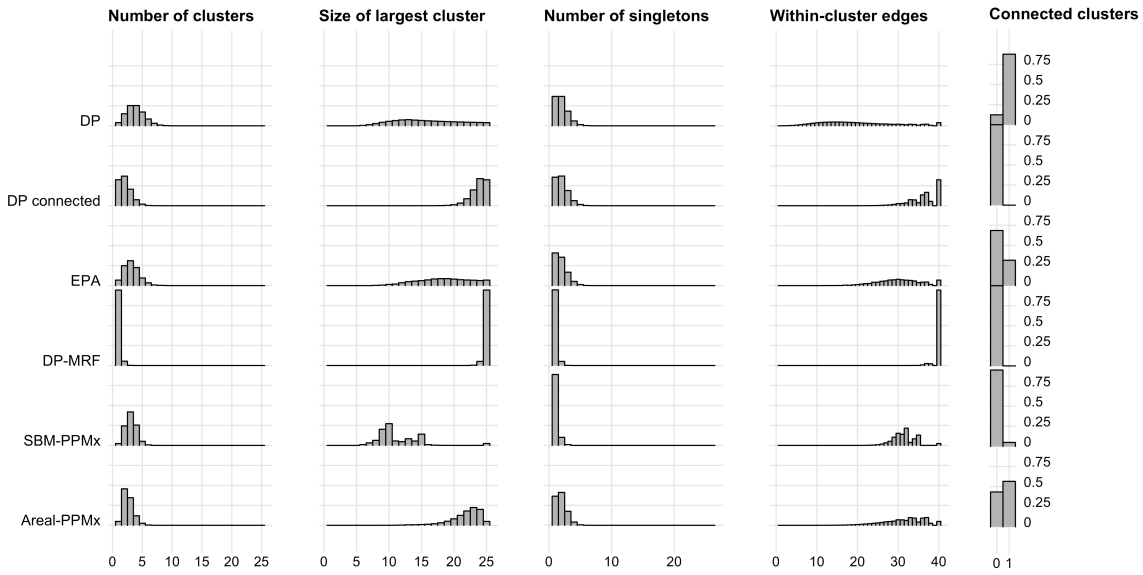


Figure 36: Distribution of the properties (a)-(e) on a 5 by 5 grid.

The prior number of failures $b$ instead can be thought as the prior number of between-cluster connections which corresponds to the parameter $c$ in the SBM:

$$b = \sqrt{n/\log(n)}.$$

## C.2. Additional prior analyses

In Figure 36 we report the graphs describing the properties (a.)-(e.) of the distributions considered for the 5 by 5 grid graph, estimated using importance sampling with a sample size of $10^8$.

# APPENDIX D

# Clustering Data at Multiple Resolutions

## D.1. Algorithm for posterior sampling

### D.1.1. Split-merge for HDP

In this section we are going to present a posterior sampling algorithm for the Hierarchical Dirichlet Process. We use the Chinese restaurant franchise representation, described by Teh et al. (2006). Instead of using the Gibbs sampling algorithm described by Teh et al. (2006), we propose a Split-Merge algorithm for the HDP, extending the work of Jain and Neal (2004b). In the Chinese Restaurant Franchise representation of the HDP, the partition is described by a partition of costumers into tables within each restaurant and a partition of tables into dishes across restaurants. Let $t_{ji}$ be the table assigned to costumer $i$ in restaurant $j$, with $\mathbf{t}_j = (t_{ji} : \forall i)$ being the partition of costumers into tables in restaurant $j$, $\mathbf{t} = (t_{ji} : \forall j, i)$ and $\mathbf{t}_{-j} = (t_{j'i} : \forall i, j' \neq j)$; moreover let $k_{jt}$ be the dish assigned to table $t$ in restaurant $j$, and $\mathbf{k} = (k_{jt} : \forall j, t)$ be the partition of tables across restaurants into dishes.

Remember that we can write the likelihood as

$$p(\mathbf{y}|\mathbf{t}, \mathbf{k}) = \prod_k \int \prod_{j,i:k_{jt_{ji}}=k} p(y_{ji}|\phi_k) d\phi_k$$

and that the prior $p(\mathbf{k}, \mathbf{t}) = p(\mathbf{k}|\mathbf{t}) \prod_j p(\mathbf{t}_j)$, where $p(\mathbf{k}|\mathbf{t})$ and $p(\mathbf{t}_j)$ are Ewens-Pitman prior distributions for partitions.

Moreover remember that for every Metropolis-Hasting proposal $\tilde{\gamma}$, we need to compute the acceptance probability $A(\tilde{\gamma}; \gamma)$ to move from partition $\gamma$ to $\tilde{\gamma}$: $A(\tilde{\gamma}; \gamma) = 1 \wedge a(\tilde{\gamma}; \gamma)$, where $a(\tilde{\gamma}; \gamma) = \frac{\pi(\tilde{\gamma}) q(\gamma; \tilde{\gamma})}{\pi(\gamma) q(\tilde{\gamma}; \gamma)}$, $q(\tilde{\gamma}; \gamma)$ the probability of proposing $\tilde{\gamma}$ from $\gamma$ and $\pi$ is the posterior distribution.

**Sampling t** We iteratively sample the partitions $\mathbf{t}_j$ for all $j$, given $\mathbf{k}$ and $\mathbf{t}_{-j}$. Two costumers $i_1$ and $i_2$ in restaurant $j$ are randomly picked and if they belong to the same cluster ($t_{ji_1} = t_{ji_2}$) a split move is performed, otherwise a merge move is implemented.

- **Split** When a split move is performed, we need to sample the new table assignment of the elements in the same cluster as $i_1$ and $i_2$. This is done similarly as Jain and Neal (2004b)'s restricted Gibbs sampling proposal. Moreover, since a new table $t_{new}$ is created, a new dish $k_{jt_{new}}$ is sampled (uniformly among the existing dishes and a new dish ). Note that this affects the partition of tables into dishes, so it needs to be taken into account in the likelihood. Let $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{k}}$ represent the split proposal for the table and the dish assignments, with probability $q(\tilde{\mathbf{t}}, \tilde{\mathbf{k}}; \mathbf{t}, \mathbf{k})$. The posterior ratio $\pi(\tilde{\mathbf{t}})/\pi(\mathbf{t})$ is given by

$$\frac{p(\{y_{ji} : \tilde{k}_{jt_{ji}} = k_1\})p(\{y_{ji} : \tilde{k}_{jt_{ji}} = k_2\})}{p(\{y_{ji} : k_{jt_{ji}} = k_1\})p(\{y_{ji} : k_{jt_{ji}} = k_2\})} \frac{\Gamma(\tilde{n}_{k_1})\Gamma(\tilde{n}_{k_2})(1 + \eta\mathbf{I}(n_{k_2} = 0))}{\Gamma(n_{k_1})\Gamma(n_{k_2})} \frac{\Gamma(\tilde{n}_{t_1})\Gamma(\tilde{n}_{t_2})\alpha}{\Gamma(n_{t_1})}$$

  Note that since the likelihood depends on the dish assignment, we need to take into account changes in the dish allocation (first term), on top of the prior for clustering costumers into tables (second term) and the one for clustering tables into dishes (third term).

- **Merge** If two tables are merged, they get assigned to the dish of $i_1$'s table and the merge happens in one unique way; however the reverse move needs to be computed. Thus, similarly to the Split move, we need to compute a *launch* split and we compute the probability to go from the launch split to the two original clusters; moreover, we compute the probability of choosing that particular dish. As before the likelihood is affected by the change in dish allocation and the prior by the change in table assignments and dish assignments.

**Sampling k** We finally sample the partition of tables into dishes. This is similarly done using a split merge algorithm which is performed in the same way as in the DP mixture

model, with the exception that now all the costumers seating at the tables corresponding to a dish are used to compute the likelihood corresponding to that cluster. Let $\tilde{\mathbf{k}}$ be the proposed dish assignment that corresponds to splitting dish $k_1$ in $\mathbf{k}$, with $k_2$ corresponding to a new dish in $\tilde{\mathbf{k}}$. In this case the posterior ratio $\pi(\tilde{\mathbf{k}})/\pi(\mathbf{k})$ is given by

$$\frac{p(\{y_{ji} : \tilde{k}_{jt_{ji}} = k_1\})p(\{y_{ji} : \tilde{k}_{jt_{ji}} = k_2\})}{p(\{y_{ji} : k_{jt_{ji}} = k_1\})} \frac{\Gamma(\tilde{n}_{k_1})\Gamma(\tilde{n}_{k_2})\eta}{\Gamma(n_{k_1})}$$

### D.1.2. Split-merge for nHDP

In the nested Hierarchical Dirichlet Process, restaurants are no longer fixed entities, but they are clusters of groups of costumers. Let $r_g$ be restaurant allocation of group $g$, and $\mathbf{r} = (r_g : \forall\, g)$. Moreover let $g_c$ be the group corresponding to costumer $c$ and let $r_{g_c}$ be the restaurant associated to costumer $c$ through its group $g$. In this model, on top of sampling $\mathbf{t}$ and $\mathbf{k}$ given the restaurant assignment, we need to sample the partition of groups into restaurants $\mathbf{r}$.

We use a Metropolis-Hastings MCMC sampling algorithm, in which the chain moves from state $x$ to state $x^*$ with probability $\alpha(x^*, x) = \min\{1, A(x^*, x)\}$ and

$$A(x^*, x) = \frac{q(x|x^*)p(x^*)}{q(x^*|x)p(x)}.$$

**Sampling t and k.** This step reduces to the split-merge sampling for the HDP described in the previous section.

**Sampling r.** Since $\mathbf{r}$ defines the division of groups into restaurants, it influences the prior probability of the assignment of costumers into tables $\mathbf{t}$. In fact, changing the number of costumers of a restaurant affects the probability of the partition, even when the clusters remain unchanged. However, changing $\mathbf{r}$ also affects the table assignment itself, because in some cases, when changing the restaurant assignment of a group, the table assignments

might become incompatible with the proposed restaurant assignment.

Consider for example the case of splitting a restaurant in which a table clusters together the costumers from two groups that are split in the two sub-restaurants. Specifically let $j_1$ and $j_2$ be these groups, such that $s = r_{j_1} = r_{j_2}$ but $s_1 = r^*_{j_1} \neq r^*_{j_2} = s_2$ and let $i_1$ and $i_2$ be two costumers such that $j_1 = g_{i_1}$ and $j_2 = g_{i_2}$. If these two costumers are sitting at the same table before the split move, $t_{si_1} = t_{si_2}$, they cannot still sit in the same table after the split move, i.e. we cannot have $t^*_{s_1 i_1} = t^*_{s_2 i_2}$, since sharing a table between two different restaurants is not possible. Thus such table assignment has probability zero given the proposed restaurant assignment and needs to be resampled together with it. That is, we need to propose, together with $\mathbf{r}^*$, a new table assignment $\mathbf{t}^*$ such that $t^*_{s_1 i_1} \neq t^*_{s_2 i_2}$.

Because of this complex requirements, our proposal is similar in spirit to a reversible-jump MCMC step. In fact, $\mathbf{r}$ affects the dimensionality of the spaces in which $\mathbf{t}$ and $\mathbf{k}$ live. As a consequence, our split and merge move for $\mathbf{r}$ is in fact a move that affects all the assignments $\mathbf{r}, \mathbf{t}$ and $\mathbf{k}$. In other words, the chain moves from state $x = (\mathbf{r}, \mathbf{t}, \mathbf{k})$ to $x^* = (\mathbf{r}^*, \mathbf{t}^*, \mathbf{k}^*)$, with a proposal that can be factorized conditionally:

$$q(x^*|x) = q(\mathbf{r}^*|\mathbf{r}) q(\mathbf{t}^*|\mathbf{r}^*, \mathbf{t}) q(\mathbf{k}^*|\mathbf{t}^*, \mathbf{k}).$$

We randomly sample two groups indices $j_1$ and $j_2$ and if $r_{j_1} = r_{j_2} = s$ we split that restaurant; if instead $s_1 = r_{j_1} \neq r_{j_2} = s_2$ we merge the restaurants $s_1$ and $s_2$.

- **Split** The split move changes the restaurant assignment of all the groups in restaurant $s$, $G_s = \{j \neq j_1, j_2 : r_j = s\}$. Let $s_1 = s$ and $s_2 = K_r + 1$ be the two sub-restaurants in $\mathbf{r}^*$, where $K_r$ is the number of clusters in $\mathbf{r}$. We assign $j_1$ to $s_1$ and $j_2$ to $s_2$, that is $r^*_{j_1} = s_1$ and $r^*_{j_2} = s_2$, and we sample $r^*_j \in \{s^*_1, s^*_2\}$ for all $j \in G_s$. This is done according to $q_{\text{split}}(\mathbf{r}^*|\mathbf{r})$.

  Given a proposed restaurant assignment $\mathbf{r}^*$, we sample the proposed table assignment

$\mathbf{t}^*$ and dish assignment $\mathbf{k}^*$, taking into account that each new restaurant $s$ in $\mathbf{r}^*$ contains all the costumers $i$ whose group $g_i$ belongs to restaurant $s$, that is all costumers $i$ such that $r^*_{g_i} = s$. Thus the table assignment $\mathbf{t}$ is changed to replace restaurant $s$ with $s_1$ and to include restaurant $s_2$. The table assignments are changed in the following way:

- if the costumers sitting at table $h$ in $\mathbf{t}_s$ all belong to groups that are assigned to $s_k$ in $\mathbf{r}^*$, the table remains unchanged in $s_k$, for $k = 1, 2$. In other words, if $\{r^*_{g_i} : t_i = h\} = \{s_1\}$, then table $h$ remains unchanged in $s_1$, and if $\{r^*_{g_i} : t_i = h\} = \{s_2\}$, then table $h$ remains unchanged in $s_2$. Moreover, the dish assignment of the table does not change: if $k_h = d$, then $k^*_h = d$.

- if the costumers sitting at table $h$ in $\mathbf{t}_s$ belong to groups that are assigned to both $s_1$ and $s_2$ in $\mathbf{r}^*$, then the table is split into two sub-tables, one for each sub-restaurant. This happens if $\{r^*_{g_i} : t_i = h\} = \{s_1, s_2\}$. The two subtables $h_1$ and $h_2$ are created deterministically, assigning to each one the costumers that belong to groups that are assigned to the corresponding sub-restaurant. So $t^*_i = h_k$ iff $r^*_{g_i} = s_k$ for $k = 1, 2$ for all $i$ such that $t_i = h$. Moreover, the dish assignment of the tables does not change: if $k_h = d$, then $k^*_{h_1} = d$ and $k^*_{h_2} = d$.

Note that these changes to the table and dish assignments, do not affect the costumer to dish assignment: even though a costumer might belong to a different restaurant or seat to a different table, its dish assignment will remain the same. Moreover, since the table and dish assignments are changed in a deterministic way, $q_{\text{split}}(\mathbf{t}^*|\mathbf{r}^*, \mathbf{t}) = 1$ and $q_{\text{split}}(\mathbf{k}^*|\mathbf{t}^*, \mathbf{k}) = 1$. Thus we only need to specify $q_{\text{split}}(\mathbf{r}^*|\mathbf{r})$. This is done with a restricted Gibbs sampling step:

$$q_{\text{split}}(r_j = s_k|\mathbf{r}_{-j}) = \frac{p(\mathbf{y}|\mathbf{k}^*, \mathbf{t}^*)p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*|\mathbf{r})n_{-j,s_k}}{\sum_{k=1,2} p(\mathbf{y}|\mathbf{k}^*, \mathbf{t}^*)p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*|\mathbf{r})n_{-j,s_k}}, \quad k = 1, 2$$

where we have denoted with $\mathbf{t}^*$ and $\mathbf{k}^*$ the table and dish assignments proposed in

the deterministic way we just described and $n_{-j,s_k}$ is the size of cluster $s_k$ excluding element $j$. Note that since the dish assignment of each costumer remains constant, we can simplify the proposal distribution above: $p(\mathbf{y}|\mathbf{k}^*, \mathbf{t}^*)$ remains constant for all $\mathbf{r}$ considered in this restricted Gibbs sampling step. Moreover, $p(\mathbf{t}^*|\mathbf{r}) \propto p(\mathbf{t}^*_{s_1}|\mathbf{r})p(\mathbf{t}^*_{s_2}|\mathbf{r})$, as the other table assignments are not affected in this step. Thus

$$q_{\text{split}}(r_j = s_k|\mathbf{r}_{-j}) = \frac{p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*_{s_1}|\mathbf{r})p(\mathbf{t}^*_{s_2}|\mathbf{r})n_{-j,s_k}}{\sum_{k=1,2} p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*_{s_1}|\mathbf{r})p(\mathbf{t}^*_{s_2}|\mathbf{r})n_{-j,s_k}}, \quad k = 1, 2$$

Note that it is also thanks to these deterministic proposal distributions that by simply sampling $\mathbf{r}$, we can sample in the multidimensional space of $(\mathbf{r}, \mathbf{t}, \mathbf{k})$. Additionally, note that to simplify the computations, instead of $p(\mathbf{k}^*|\mathbf{t}^*)$ we consider $p(\mathbf{k}^*|\mathbf{t}^*)/p(\mathbf{k}|\mathbf{t})$. This can simply be computed as $\prod_{k \in D_s} \Gamma(n_k + m_k)/\Gamma(n_k)$, where $n_k$ is the number of tables belonging to dish $k$ in the original assignment $\mathbf{k}$, $m_k$ is how many of those tables were split into two sub-tables, and $D_s$ is the set of dishes served in restaurant $s$.

When low resolution data is also available, we can simplify our proposal distribution $q_{\text{split}}$ and only base it on the likelihood of $\mathbf{y}^{(L)} = (y_j)_j$:

$$q_{\text{split}}(r_j = s_k|\mathbf{r}_{-j}) = \frac{p(\mathbf{y}^{(L)}|\mathbf{r})n_{-j,s_k}}{\sum_{k=1,2} p(\mathbf{y}^{(L)}|\mathbf{r})n_{-j,s_k}}, \quad k = 1, 2$$

- **Merge** The merge move changes the restaurant assignment of all the groups in restaurants $s_1$ and $s_2$, $G_{s_1,s_2} = \{j : r_j \in \{s_1, s_2\}\}$. Let $s$ be the new restaurant which will replace $s_1$ and let $r^*_j = s$ for all $j \in G_{s_1,s_2}$ (restaurant $s_2$ gets removed from $\mathbf{r}^*$). Note that $q_{\text{merge}}(\mathbf{r}^*|\mathbf{r}) = 1$.

  As the restaurant assignment is changed, the table and dish assignments need to change too. As before, we need to take into account that a new restaurant $s$ in $\mathbf{r}^*$ contains all the costumers $i$ whose group $g_i$ belongs to restaurant $s$, that is $r^*_{g_i} = s$. A

naive proposal for changing the table assignment would be to move all tables of $s_2$ to the new merged restaurant $s$. However, we need to choose a proposal that can make the split move reversible. For this reason, in the table assignment proposal we need to merge some tables that belonged to the two restaurants.

Specifically, for each dish $d$ we consider the tables in the two restaurants $s_1$ and $s_2$ that were assigned to dish $d$. Let $T^d_{s_k} = \{h \text{ table in } s_k : k_h = d\}$ for $k = 1, 2$. If there is at least one such table in each restaurant, i.e. $\#T^d_{s_k} > 0$ for both $k = 1, 2$, we combine tables into pairs. This is done by considering the restaurant with the least number of such tables, say $s_1$, and considering a one-to-one function $f$ from its tables $T^d_{s_1}$ to the ones in the other restaurant $T^d_{s_2}$, sampled uniformly at random. Thus, if $k_d = \#T^d_{s_1}$ and $n_d = \#T^d_{s_2}$, the probability of sampling $f$ is $\frac{1}{n_d!/(n_d-k_d)!}$. Given this matching $f$, we consider the probability of merging or not the tables in each pair. We do this proportionally to the probability of the partitions of tables in the merged restaurant $p(\mathbf{t}^*_s)$. The probability of merging a given pair of tables $(h_1, h_2)$ is 0.5. Note that if tables $h_1$ and $h_2$ are merged in table $h$, then $t^*_i = h$ for all $i$ such that $t_i = h_k$, for $k = 1, 2$; otherwise $t^*_i = t_i = h_k$. Note that in either case the dish assignment will not change, $k^*_h = d$ in the former case, or $k^*_{h_k} = d$ for $k = 1, 2$ in the latter.

Thus the overall probability of the new table assignment $\mathbf{t}^*$ is given by

$$q_{\text{merge}}(\mathbf{t}^*|\mathbf{r}^*, \mathbf{t}) = \prod_d \left[ \frac{1}{n_d!/(n_d - k_d)!} \prod_{i=1}^{k_d} \left( [p^m_{t_1,t_2}]^{\mathbf{I}(\text{m})} [1 - p^m_{t_1,t_2}]^{\mathbf{I}(\text{s})} \right) \right]$$

where $\mathbf{I}(m)$ and $\mathbf{I}(s)$ are the indicators of a split or a merge. Note that, as in the split move, the choice for $\mathbf{k}^*$ is deterministic and $q_{\text{merge}}(\mathbf{k}^*|\mathbf{t}^*, \mathbf{k}) = 1$.

Remember now that to find the acceptance probability $\alpha(x^*, x)$ we need to consider the

proposed move and the reverse move. Thus, to compute $A_{\text{split}}(x^*, x)$ we have

$$\begin{aligned} A_{\text{split}}(x^*, x) &= \frac{q_{\text{merge}}(x|x^*)p(x)}{q_{\text{split}}(x^*|x)p(x)} = \\ &= \frac{q_{\text{merge}}(\mathbf{t}|\mathbf{r}, \mathbf{t}^*)}{q_{\text{split}}(\mathbf{r}^*|\mathbf{r})} \frac{p(\mathbf{y}|\mathbf{t}, \mathbf{k})p(\mathbf{k}|\mathbf{t})p(\mathbf{t}|\mathbf{r})p(\mathbf{r})}{p(\mathbf{y}|\mathbf{t}^*, \mathbf{k}^*)p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*|\mathbf{r}^*)p(\mathbf{r}^*)} = \\ &= \frac{q_{\text{merge}}(\mathbf{t}|\mathbf{r}, \mathbf{t}^*)}{q_{\text{split}}(\mathbf{r}^*|\mathbf{r})} \frac{p(\mathbf{k}|\mathbf{t})p(\mathbf{t}|\mathbf{r})p(\mathbf{r})}{p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*|\mathbf{r}^*)p(\mathbf{r}^*)}, \end{aligned}$$

where the likelihood ratio $p(\mathbf{y}|\mathbf{t}, \mathbf{k})/p(\mathbf{y}|\mathbf{t}^*, \mathbf{k}^*)$ can be ignored because it's equal to 1. Similarly, for $A_{\text{merge}}(x^*, x)$ we have

$$\begin{aligned} A_{\text{merge}}(x^*, x) &= \frac{q_{\text{split}}(x|x^*)p(x)}{q_{\text{merge}}(x^*|x)p(x)} = \\ &= \frac{q_{\text{split}}(\mathbf{r}|\mathbf{r}^*)}{q_{\text{merge}}(\mathbf{t}^*|\mathbf{t}, \mathbf{r}^*)} \frac{p(\mathbf{k}|\mathbf{t})p(\mathbf{t}|\mathbf{r})p(\mathbf{r})}{p(\mathbf{k}^*|\mathbf{t}^*)p(\mathbf{t}^*|\mathbf{r}^*)p(\mathbf{r}^*)}. \end{aligned}$$

# LIST OF TABLES

# LIST OF ILLUSTRATIONS

# BIBLIOGRAPHY


P. Agrawal, L. S. Tekumalla, and I. Bhattacharya. Nested hierarchical dirichlet process for nonparametric entity-topic analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 564–579. Springer, 2013.

S. Aldor-Noiman, L. D. Brown, E. B. Fox, and R. A. Stine. Spatio-temporal low count processes with application to violent crime events. *Statistica Sinica*, pages 1587–1610, 2016.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII1983*, pages 1–198. Springer, 1985.

C. Anderson and L. M. Ryan. A comparison of spatio-temporal disease mapping approaches including an application to ischaemic heart disease in new south wales, australia. *International Journal of Environmental Research and Public Health*, 14:146, 2017.

C. Anderson, D. Lee, and N. Dean. Spatial clustering of average risks and risk trends in bayesian disease mapping. *Biometrical Journal*, 59(1):41–56, 2017.

L. Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 2013.

C. Balocchi and S. T. Jensen. Spatial modeling of trends in crime over time in Philadelphia. *Annals of Applied Statistics*, 13(4):2235–2259, 2019.

C. Balocchi, S. K. Deshpande, E. I. George, and S. T. Jensen. Bayesian spatial clustering of crime in Philadelphia with particle optimization. *arXiv preprint arXiv:1912.00111 [stat.AP]*, 2019.

S. Banerjee, B. P. Carlin, P. Li, and A. M. McBean. Bayesian areal wombling using false discovery rates. *Statistics and its Interface*, 5(2):149–158, 2012.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

L. Bernardinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini. Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, 14 (21-22):2433–2443, 1995.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in

spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991. ISSN 00203157. doi: 10.1007/BF00116466.

N. Best, S. Richardson, and A. Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical methods in medical research*, 14(1):35–59, 2005.

D. Blackwell, J. B. MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57 (2):7, 2010.

B. Boots. Using local statistics for boundary characterization. *GeoJournal*, 53(4):339–345, 2001.

P. J. Brantingham and P. L. Brantingham. *Patterns in crime*. Macmillan New York, 1984.

D. Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51(3/4):481–483, 1964.

J. B. Burbidge, L. Magee, and A. L. Robb. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83 (401):123–127, 1988a.

J. B. Burbidge, L. Magee, and A. L. Robb. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83 (401):123–127, 1988b. ISSN 01621459. URL `http://www.jstor.org/stable/2288929`.

F. Camerlenghi, D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez. Latent nested nonparametric priors. *arXiv preprint arXiv:1801.05048*, 2018.

G. Casella, E. Moreno, and F. J. Girón. Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613 – 685, 2004.

G. Casella, E. Moreno, F. J. Girón, et al. Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613–658, 2014.

G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451): 957 – 970, 2000.

H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

N. Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

D. B. Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4 (2):243 – 264, 2009.

D. B. Dahl, R. Day, and J. W. Tsai. Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732, 2017.

D. Denison and C. Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149, 2001.

F. Denti. *Bayesian Mixtures for Large Scale Inference*. PhD thesis, Università degli Studi di Milano-Bicocca, 2020.

M. M. Ehlenz. Neighborhood revitalization and the anchor institution: Assessing the impact of the university of pennsylvania's west philadelphia initiatives on university city. *Urban Affairs Review*, 52(5):714 – 750, 2016.

W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.

FBI. Uniform Crime Reporting program, definitions. `https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offense-definitions`.

W. Feng, C. Y. Lim, T. Maiti, and Z. Zhang. Spatial regression and estimation of disease risks: A clustering-based approach. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(6):417–434, 2016.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

S. R. Flaxman. A general approach to prediction and forecasting crime rates with gaussian processes. *Heinz College Technical Report*, 2014.

O. François, S. Ancelet, and G. Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2):805–816, 2006.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881 – 889, 1993.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. J. Green and S. Richardson. Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070, 2002.

J. A. Hartigan. Partition models. *Communications in statistics-Theory and methods*, 19 (8):2745–2756, 1990.

D. T. Herbert. *The geography of urban crime*. Longman London, 1982.

J. A. Hoeting, D. Madigan, A. E. Raferty, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382 – 417, 1999.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

M. Hosseinpouri and M. J. Khaledi. An area-specific stick breaking process for spatial data. *Statistical Papers*, 60(1):199–221, 2019.

C. Humphrey, S. T. Jensen, D. S. Small, and R. Thurston. Analysis of urban vibrancy and safety in philadelphia. Working Paper. Arxiv: 1702.07909, 2017.

M. Hvistendahl. Crime forecasters. *Science*, 353:1484–1487, 2016.

S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158 – 182, 2004a. doi: 10.1198/1061860043001.

S. Jain and R. M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of computational and Graphical Statistics*, 13(1):158–182, 2004b.

S. Jo, J. Lee, P. Müller, F. A. Quintana, L. Trippa, et al. Dependent species sampling models for spatial density estimation. *Bayesian Analysis*, 12(2):379–406, 2017.

S. D. Johnson, K. J. Bowers, et al. Stable and fluid hotspots of crime: Differentiation and identification. *Built Environment*, 34(1):32–45, 2008.

T. D. Johnson, Z. Liu, A. J. Bartsch, and T. E. Nichols. A bayesian non-parametric potts model with application to pre-surgical fmri data. *Statistical methods in medical research*, 22(4):364–381, 2013.

D. Kahle and H. Wickham. ggmap: spatial visualization with ggplot2. *The R Journal*, 5 (1), 2013.

L. Knorr-Held and G. Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

J. W. Lau and P. J. Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.

J. Law, M. Quick, and P. Chan. Bayesian spatio-temporal modeling for analysing local

patterns of crime over time at the small-area level. *Journal of quantitative criminology*, 30(1):57–78, 2014.

A. B. Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press, 2013.

D. Lee. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89, 2011. ISSN 18775845. doi: 10.1016/j.sste.2011.03.001. URL `http://dx.doi.org/10.1016/j.sste.2011.03.001`.

D. Lee and R. Mitchell. Boundary detection in disease mapping studies. *Biostatistics*, 13 (3):415–426, 2012.

D. Lee and R. Mitchell. Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4):593–608, 2013.

B. G. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer, 2000.

J. LeSage and R. K. Pace. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.

G. Li, R. Haining, S. Richardson, and N. Best. Space–time variability in burglary risk: a bayesian spatio-temporal modelling approach. *Spatial Statistics*, 9:180–191, 2014.

P. Li, S. Banerjee, and A. M. McBean. Mining boundary effects in areally referenced spatial data using the bayesian information criterion. *Geoinformatica*, 15(3):435–454, 2011.

H. Lu and B. P. Carlin. Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37(3):265–285, 2005.

H. Lu, C. S. Reilly, S. Banerjee, and B. P. Carlin. Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452, 2007.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

J. Møller and R. P. Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

P. A. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

R. Neal. Markov Chain sampling methods for Dirichlet Process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249 – 265, 2000.

P. Orbanz and J. M. Buhmann. Nonparametric bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45, 2008.

D. T. OBrien and C. Winship. The gains of greater granularity: The presence and persistence of problem properties in urban neighborhoods. *Journal of quantitative criminology*, 33(3):649–674, 2017.

G. L. Page, F. A. Quintana, et al. Spatial product partition models. *Bayesian Analysis*, 11 (1):265–298, 2016.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2014.

J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226, 2010.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

J. Pitman. Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

J. Pitman et al. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for , 2002.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179 – 191, 1997.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

L. Ren, L. Du, L. Carin, and D. Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(Jan):203–239, 2011.

V. Ročková. Particle EM for variable selection. *Journal of the American Statistical Association*, 113(524):1684 – 1697, 2018.

A. Rodríguez, D. B. Dunson, and A. E. Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

C. Schnell, A. A. Braga, and E. L. Piza. The influence of community areas, neighborhood

clusters, and street segments on the spatial variability of violent crime in chicago. *Journal of quantitative criminology*, 33(3):469–496, 2017.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

A. Singh, G. Nadkarni, J. Guttag, and E. Bottinger. Leveraging hierarchy in medical codes for predictive modeling. In *Proceedings of the 5th ACM conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 96–103. ACM, 2014.

W. Steenbeek and D. Weisburd. Where the action is in crime? an examination of variability of crime across different spatial units in the hague, 2001–2009. *Journal of Quantitative Criminology*, 32(3):449–469, 2016.

M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

M. A. Taddy. Autoregressive mixture models for dynamic spatial poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association*, 105(492):1403–1417, 2010.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302.

L. S. Tekumalla, P. Agrawal, and I. Bhattacharya. Nested hierarchical dirichlet processes for multi-level non-parametric admixture modeling. *arXiv preprint arXiv:1508.06446*, 2015.

W. Tobler. On the first law of geography: A reply. *Annals of the Association of American Geographers*, 94(2):304–310, 2004.

S. Wade, Z. Ghahramani, et al. Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, 2017.

S. Wade, Z. Ghahramani, et al. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

C. Wang and D. M. Blei. A split-merge mcmc algorithm for the hierarchical dirichlet process. *arXiv preprint arXiv:1201.1657*, 2012.

P. Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

W. H. Womble. Differential systematics. *Science*, 114(2961):315–322, 1951. ISSN 0036-8075. doi: 10.1126/science.114.2961.315. URL `http://science.sciencemag.org/content/114/2961/315`.

Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015.

Y. Zenou. The spatial aspects of crime. *Journal of the European Economic Association*, 1 (2-3):459–467, 2003.