Burglary in London: Insights from Statistical Heterogeneous Spatial Point Processes

Jan Povala

Department of Mathematics, Imperial College London; The Alan Turing Institute

E-mail: jan.povala11@imperial.ac.uk

Seppo Virtanen, Mark Girolami

Department of Engineering, University of Cambridge; The Alan Turing Institute

[Accepted at the Journal of the Royal Statistical Society. Series C: Applied Statistics]
arXiv: 1910.05212

This version: June 12 2020
First version: October 9 2019

Abstract. To obtain operational insights regarding the crime of burglary in London, we consider the estimation of the effects of covariates on the intensity of spatial point patterns. Inspired by localised properties of criminal behaviour, we propose a spatial extension to mixtures of generalised linear models from the mixture modelling literature. The proposed Bayesian model is a finite mixture of Poisson generalised linear models such that each location is probabilistically assigned to one of the groups. Each group is characterised by the regression coefficients, which we subsequently use to interpret the localised effects of the covariates. By using a blocks structure of the study region, our approach allows specifying spatial dependence between nearby locations. We estimate the proposed model using Markov Chain Monte Carlo methods and provide a Python implementation.

1. Introduction

Use of statistical models for understanding and predicting criminal behaviour has become increasingly relevant for police forces, and policymakers (Felson & Clarke 1998, Bowers & Hirschfield 1999, PredPol 2019). While short-term forecasting of criminal activity has been used to allocate policing resources better (Taddy 2010, Mohler et al. 2011, Aldor-Noiman et al. 2017, Flaxman et al. 2019, PredPol 2019), understanding the criminal behaviour and target selection process through statistical models has a potential to be used for designing policy changes and development programs (Felson & Clarke 1998). In this work, we consider the problem of burglary crime in London. In the UK, burglary is a well-reported crime, but the detection rate remains at the 10-15% level (Smith et al. 2013). Rather than being concerned with short-term forecasting, we focus on understanding the effects of spatially varying explanatory variables on the target selection through descriptive regression models. Inferences made using these models help us understand the underlying mechanisms of burglary. The main contribution of this work is the integration of statistical methods in spatial modelling with the findings from the criminological literature.

Instances of burglary can be represented as a *spatial point pattern* – a finite or countably infinite set of points in the study region. Understanding the intensity of the occurrences through spatially varying covariates is the main objective of this work. The task of estimating the effects of the covariates on the intensity can be classified as a multivariate regression modelling, in which systematic effects of the explanatory variables are of interest while taking into account other random effects such as measurement errors and spatial correlation (McCullagh & Nelder 1998). In the context of spatial data, it has been widely recognised that multivariate regression modelling techniques which do not account for *spatial dependence* and *spatial heterogeneity* can lead to biased results and faulty inferences (Anselin et al. 2000). Spatial dependence refers to the

Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things"(Tobler 1970). Spatial dependence manifests mostly in the spatial correlation of the residuals of a model. In non-spatial settings, the residuals are often assumed to be independent and identically distributed (McCullagh & Nelder 1998). Spatial heterogeneity is exhibited when the object of interest, in our case, the intensity of a point pattern, shows location-specific behaviour. For example, properties of the burglary point pattern in a city centre are going to be different from the properties in a residential area. Formalising these two concepts and incorporating them into modelling methodology results in more accurate spatial models (Anselin et al. 2000).

Log-Gaussian Cox process (Møller et al. 1998, Møller & Waagepetersen 2007) has been a common approach for modelling intensity of spatial point patterns (Diggle et al. 2013, Serra et al. 2014, Flaxman et al. 2015). The flexibility of the model is due to the Gaussian process part through which complex covariance structures, including spatial dependence and heterogeneity, can be accounted for. In practice, stationary covariance functions are used for computational reasons (Diggle et al. 2013). As a result, log-Gaussian Cox process models with stationary covariance functions handle spatial dependence but do not account for spatial heterogeneity.

Mixture based approaches have been adopted as a way of enriching the collection of probability distributions to account for spatial heterogeneity often observed in practice (Green 2010, Fernández & Green 2002). Notably, Knorr-Held & Raßer (2000), Fernández & Green (2002), Green & Richardson (2002) used mixtures for modelling the elevations of disease prevalence. While these methods improve the model fit by accounting for spatial heterogeneity as wells as spatial dependence, they provide little interpretation as to why the level is elevated in certain areas. Also, these three methods have been tested only at a modest scale. Following this line of work, Hildeman et al. (2018) proposed a method in which each mixture component can take a rich representation that may include covariates. Although this model is very rich in representation, the empirical study in the paper was limited to the case of two mixtures, with one of the components being held constant. Their study of a tree point pattern and its dependence on soil type was carried out on a region discretised into a grid with 2461 cells.

A very different approach to controlling for spatial heterogeneity has been taken by Gelfand et al. (2003) who allow regression coefficients to vary across the spatial region. The method treats the coefficients of the covariates as a multivariate spatial process. The process is, however, very challenging to fit and is often limited to 2 or 3 covariates (Banerjee et al. 2015, p.288). A simpler version of the same idea is geographically-weighted regression (Brunsdon et al. 1996), where the regression coefficients are weighted by a latent component whose properties have to be specified a priori or learned through cross-validation.

Motivated by the computational challenges and limited interpretability of the aforementioned approaches, we propose a mixture based method that takes into account spatial dependence and is able to discover latent groups of locations and characterise each group by group-specific effects of spatially varying covariates. To estimate the model parameters from the limited data and to quantify the uncertainty of the estimates, we follow the Bayesian framework.

More specifically, our approach builds upon the mixtures of generalised linear models (Grün & Leisch 2008), in which observations are modelled as a mixture of different models. We cater for spatial dependence using an approach inspired by Fernández & Green (2002) and Knorr-Held & Raßer (2000). Our model probabilistically assigns each location to a particular mixture component, while imposing spatial dependence through prior information. The prior information will suggest that locations that are close to each other are likely to belong to the same component. We define a pair of locations to be close if both of them are in the same block. We use the blocking structure predefined by the census tracts, but our method allows defining custom ones. We further model spatial dependence of the blocks using latent Gaussian processes, following Fernández & Green (2002). The posterior inferences for the individual components consisting of regressions coefficients and the assignments of locations to clusters are used to draw conclusions and provide insights about the heterogeneity of the spatial point pattern across the study region.

In contrast to Fernández & Green (2002) and Green & Richardson (2002), this work considers including the covariates into each mixture component, rather than having intercept-only components. Compared to the approach of Hildeman et al. (2018) who model the log-intensity of a point pattern as a mixture of Gaussian random fields, our model is more constrained but provides better scalability.

We show that the proposed methodology effectively models burglary crime in London. By comparing our approach to log-Gaussian Cox process (LGCP), a standard model for spatial point patterns (Diggle et al. 2013), we show that our method outperforms LGCP and is more computationally tractable. Lastly, the interpretation of inferred quantities provides useful criminological insights.

The rest of the paper is structured as follows. Section 2 defines the model and details the inference method, section 3 elaborates on our application and gives the discussion of model choices specific to our application. The obtained results are discussed in section 4. Section 5 concludes the paper.

2. Modelling methodology

It is widely recognised that burglary crime is spatially concentrated (Brantingham & Brantingham 1981, Clare et al. 2009, Johnson & Bowers 2010). It is also apparent that some areas in the study region will exhibit extreme behaviour. For example, areas with no buildings such as parks will have no burglary for structural reasons. To effectively model burglary, these phenomena need to be accounted for using *spatial effects*. The two important spatial effects are *spatial dependence* and *spatial heterogeneity* (Anselin et al. 2000).

For our modelling framework, we choose the Bayesian paradigm because it allows us to formalise prior knowledge, and to quantify uncertainty in the unknown quantities of our model. In our application, burglary data are given as a point pattern over a fixed period of time. We discretise the point pattern onto a grid of N cells by counting the points in each cell. Although any form of discretisation is allowed, throughout this paper, we work with a regular grid.

We model the count of points in a cell n, y_n, conditioned on the mixture component k as a Poisson-distributed random variable, with the logarithm of the intensity driven by a linear term, which is specific for each mixture component, indexed by $k = 1, \dots, K$. The linear term is a linear combination of J covariates for cell n, $\boldsymbol{X}_n$, and the corresponding coefficients, $\boldsymbol{\beta}_k$. The covariates need to be specified for the application of interest and usually include the intercept. To specify the prior distribution for the regression coefficients, we use a prior that shrinks the estimate towards zero. For each coefficient, we set $\beta_{k,j} \sim \mathcal{N}(0, \sigma^2_{k,j})$, where $\sigma^2_{k,j} \sim \mathrm{InvGamma}(1, 0.01)$. We put the uniform prior on the intercepts, if present.

Each cell n is probabilistically allocated to one of the K components through an allocation variable, z_n, which is a categorical random variable with event probabilities given by the mixture weights prior, $\boldsymbol{\pi}_{b[n]}$. The value of $\boldsymbol{\pi}_{b[n]}$ is shared for all locations within cell n's block, $b[n]$. The blocks for the study region are defined as non-overlapping spatial areas spanning the whole study region. In many practical applications, the block structure is already defined by administrative units or census tracts. Block $b[n]$ is the block that contains the centroid point of cell n. The block-specific event probabilities will express the belief that the effect of the covariates is the same within the block unless evidence from the observed data outweighs this information.

To model the mixture weights prior for block b, $\boldsymbol{\pi}_b = (\pi_{1,b}, \dots, \pi_{K,b})$, we allow for different choices provided that $\pi_{k,b} \geq 0$ and $\sum_{k=1}^{K} \pi_{k,b} = 1$, i.e. it is a valid probability measure. One possible choice which also takes into account the spatial dependence between the blocks is to model the mixture weights prior for block b and mixture component k as

$$\pi_{k,b} = \frac{\exp(\mathrm{f}_{k,b})}{\sum_{l=1}^{K} \exp(\mathrm{f}_{l,b})},$$

where $\mathrm{f}_{k,b}$ is the evaluation of f_k at the centroid of block b and f_k is an independent zero-mean Gaussian Process with hyper-parameters $\boldsymbol{\theta}_k$. The prior for $\boldsymbol{\theta}_k$ is specified depending on the kernel function used. We will use the squared exponential kernel throughout this work (Rasmussen & Williams 2006).

We refer to the proposed model as a *spatially-aware mixture of Poisson generalised linear models* (SAM-GLM). The formulation is summarised in the equation and the graphical representation shown in figure 1. In the proposed model, we handle *spatial heterogeneity* using the mixture components, each of which specifies a set of J regression coefficients, $\boldsymbol{\beta}_k$. *Spatial dependence* is considered first within each block and also through inter-block dependence imposed by K Gaussian processes. Modelling the spatial dependence using Gaussian processes at the block level instead of cell level allows for more efficient estimation procedures as we discuss later.

$$\begin{aligned}
\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{X}_n &\sim \text{Poisson}\left(\exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k\right)\right) \\
\mathrm{z}_n|\boldsymbol{\pi} &\sim \text{Cat}(\pi_{1,b[n]}, \ldots, \pi_{K,b[n]}) \\
\boldsymbol{\pi}_{k,b}|f_k &= \frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K} \exp(\mathrm{f}_{l,b[n]})} \\
f_k|\boldsymbol{\theta}_k &\sim \mathcal{GP}(0, \kappa_{\boldsymbol{\theta}_k}(\cdot,\cdot)) \\
\boldsymbol{\theta}_k &\sim \text{kernel-dependent prior} \\
\beta_{k,j}|\sigma^2_{k,j} &\sim \mathcal{N}(0, \sigma^2_{k,j}) \\
\sigma^2_{k,j} &\sim \text{InvGamma}(1, 0.01).
\end{aligned}$$

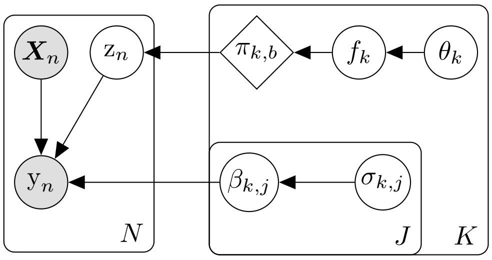


Figure 1. Summary of the SAM-GLM model and its graphical representation.

2.1. Excess of zeros, overdispersion

Two common challenges encountered when modelling count data using standard Poisson generalised linear models (GLM) are *excess of zeros* and *overdispersion* (McCullagh & Nelder 1998, Breslow 1984). The former refers to the presence of zeros that are structural, rather than due to chance. In the context of burglary, structural zeros occur in locations with no buildings, e.g. parks. The latter issue refers to the situation when the variability of the observed data is higher than what would be expected based on a particular statistical model. The standard Poisson GLM for the burglary point pattern, a special case of our model ($K = 1$), suffers from overdispersion for different specifications of the covariates term – see section A in the appendix. The flexibility of our proposed model can account for the excess of zeros by identifying a low-count component to which areas of low intensity will be assigned. Similarly, introducing mixtures can reduce overdispersion. Two cells with similar values for the covariates, but with very different observed counts are likely to have the same expected count under the standard Poisson GLM. Under the mixture model, each cell would be allowed to follow a different model.

2.2. Inference

Statistical inference in the Bayesian setting involves inferring the posterior probability distribution for the quantities of interest. In this work, we choose the Markov Chain Monte Carlo (MCMC) method to sample from the posterior distributions (Gelman et al. 2013).

Firstly, the scale parameter for the regression coefficients, σ^2_{kj}, is analytically integrated out to simplify the inference (see equation 23 in the appendix). The quantities of interest are the allocation vector $\mathbf{z}$, regression coefficient vector for each mixture component, $\boldsymbol{\beta}_k$, unnormalised mixture weights priors at the centroids of the blocks, $\mathrm{f}_{k,b}$, and its hyper-parameters. For brevity, let $\boldsymbol{\beta}$ be a $K \times J$ matrix of regression coefficients for all mixture components and each covariate, $\boldsymbol{X}$ be an $N \times J$ matrix of all covariates for each location, $\boldsymbol{F}$ be a $B \times K$ matrix such that $\boldsymbol{F}_{b,k} = \mathrm{f}_{k,b}$, and $\boldsymbol{\theta}$ the vector of kernel hyperparameters for all f_k's. The unnormalised joint posterior probability distribution is given as

$$p(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{F}, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{X}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z})p(\mathbf{z}|\boldsymbol{F})p(\boldsymbol{F}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\beta}) \quad (1)$$

We employ the Metropolis-within-Gibbs scheme (Geman & Geman 1984, Metropolis et al. 1953) and sample from the posterior in three steps:

(a) We sample the regression coefficients $\beta_{k,j}$ jointly for all $k = 1, \ldots, K$ and $j = 1, \ldots, J$. The unnormalised density of the conditional distribution is given as

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \mathbf{y}, \mathbf{z}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z})p(\boldsymbol{\beta}). \tag{2}$$

Equation 2 is sampled using the Hamiltonian Monte Carlo method (Duane et al. 1987), for which efficient sampling schemes are available, e.g. Girolami & Calderhead (2011).

(b) Mixture allocation is sampled cell by cell directly using the following equation

$$p(\mathrm{z}_n = k|\mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}, \boldsymbol{F}) \propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k)\frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K}\exp(\mathrm{f}_{l,b[n]})} \tag{3}$$

(c) We sample all K functions with the GP prior and their hyperparameters jointly using the Hamiltonian Monte Carlo. The joint posterior density is proportional to the expression below

$$p(\boldsymbol{F}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K}\exp(\mathrm{f}_{l,b[n]})}\right)^{I(z_n=k)}\prod_{k=1}^{K}p(f_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k), \tag{4}$$

where $I(\cdot)$ is the indicator function.

For the full expansion of the conditional distributions in equations (2), (3), and (4), see section C in the appendix.

In terms of computational tractability, equation 2 takes $\mathcal{O}(N + J)$ steps, equation 3 requires $\mathcal{O}(N \times K)$ steps, and equation 4 requires $\mathcal{O}\big(B^3 \times K\big)$ steps due to matrix inversions of size $B \times B$ for each of the K components. To contrast it with a standard model for spatial point patterns, one sample from a log-Gaussian Cox process involves matrix inversions that require $\mathcal{O}\big(N^3\big)$ steps (Diggle et al. 2013). Thanks to blocking, the inference requires inversions of smaller matrices.

2.3. Special case: independent blocks

The model and the associated inference introduced in this section provide a very flexible framework for modelling the spatial dependence of cells via blocks that are also spatially dependent. However, this comes at a high cost – inferring posterior distribution over K Gaussian processes that are combined using the logistic function is challenging at scale as each sample requires $\mathcal{O}\big(B^3 \times K\big)$ operations.

If we assume that the mixture weights priors ($\boldsymbol{\pi}_b$) for all blocks are independent and, conditioned on α, distributed as

$$\boldsymbol{\pi}_b|\alpha \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \tag{5}$$

the inference becomes more tractable. Specifically, equation 4 is not needed anymore, equation 2 stays the same, and equation 3 is replaced by

$$p(\mathrm{z}_n = k|\mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n\boldsymbol{\beta}, \mathbf{y}) \propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n\boldsymbol{\beta}_k)\frac{c^{\bar{n}}_{b[n]k} + \alpha}{K\alpha + \sum_{i=1}^{K} c^{\bar{n}}_{b[n]k}}. \tag{6}$$

As a result, the time complexity to take one sample from the unknown quantities is dominated by resampling z_n's in equation 6, which can be computed in $\mathcal{O}(N \times K)$ steps. For the full derivation of equation 6, see section C.2.4 in the appendix.

In the literature, $\alpha = 1/K$ is a recommended choice, see, e.g., Alvares et al. (2018). This prior formulation induces sparsity and is able to cancel out components in an overfitted mixture (Rousseau & Mengersen 2011). In the experiments we compare the trade-off between computational complexity and modelling flexibility.

2.4. Identifiability

Specifying a mixture model means that the model likelihood is invariant under the relabelling of the mixture components (Celeux et al. 2000). This issue is commonly referred to as lack of identifiability. In the context of SAM-GLM model, $p(\mathbf{y}|\mathbf{z}, \boldsymbol{X}, \boldsymbol{\beta})$ is invariant under the relabelling of $\boldsymbol{\beta}_k$ and f_k's, which are the component-specific model parameters.

Exploration in high dimensional spaces is in general hard for an MCMC sampler. As the dimension of the parameter space for the mixture model increases, the sampler is likely to explore only one of the $K!$ possible modes. For the sampler to switch to a different mode, it would have to get past the area of low probability mass surrounding the chosen mode. However, note that as the number of mixture components increases, the chance of the sampler switching to a different mode increases as the shortest distance between a pair of component-specific parameters is likely to decrease.

Since the identifiability issue poses a problem only for the interpretation of the parameters, we inspect the traceplot of the Markov chain for each identifiable parameter to assert that relabelling is not present when interpreting the mixtures.

3. Application: London burglary crime

3.1. Data description

The methodology above has been developed to enable the analysis of our application − burglary in London. The data, published online by the UK police forces (police.uk 2019), are provided monthly as a spatial point pattern over the area of $1572\,\mathrm{km}^2$ of both residential and non-residential burglary occurrences. The non-residential burglary refers to instances where the target is not a dwelling, e.g., commercial or community properties. We discretise our study area into a regular grid by counting the number of burglary occurrences within each cell. We choose a grid for computational reasons when comparing to competing methods (see section B in the appendix). Given our focus on spatial modelling, we temporally aggregate the point pattern into two datasets: the one-year dataset, starting 01/2015 and ending 12/2015, with 70,234 burglaries, and the three-year dataset, starting 01/2013 and ending 12/2015, with 224,747 burglaries.

Our analysis uses land use data, socioeconomic census data from 2011, and points of interest data from 2018 to estimate their effect on the intensity of the burglary point pattern. Land use data are available as exact geometrical shapes. The census variables are measured with respect to census tracts, called output areas (OA). The OAs have been designed to have similar population sizes and be as socially homogeneous as possible, based on the tenure of households and dwelling types. Each of the 25,053 OAs in London has between 100 people or 40 households and 625 people or 250 households. The OAs are aggregated into 4,835 lower super output areas (LSOA), which in turn are aggregated into 983 middle super output areas (MSOA). An LSOA has at least 1,000 people or 400 households and at most 3,000 people or 1,200 households. For an MSOA, the minimum is 5,000 people or 2,000 households, and the maximum is 15,000 people or 6,000 households. The points of interest data are given as a point pattern. To project the data measured at non-grid geometries (the census and land use data) onto the grid we use weighted interpolation. The method assumes that the data is uniformly distributed across the OA. For cells that have an overlap with more than one OA, we compute the value for each such cell by combining the overlapping OAs and adjusting for the size of the overlap.

3.2. Criminology background

We use existing criminology studies to identify explanatory variables and formulate hypotheses about burglary target selection. The target choice is a decision-making process of maximising *reward* with minimum *effort*, and managing the *risk* of being caught (a process analogous to optimal foragers in wildlife (Johnson & Bowers 2004)). Therefore, we categorise the explanatory variables into these three categories: reward, effort, and risk.

3.2.1. Reward, opportunities, attractiveness

Theoretically supported by rational choice theory (Clarke & Cornish 1985), offenders seek to maximise their reward by choosing areas of many opportunities and attractive targets. Firstly, the *number of dwellings* is used in the literature as a measure of the abundance of residential targets (Bernasco & Nieuwbeerta 2005, Clare et al. 2009, Townsley et al. 2015, 2016). *Real estate prices* and *household income* have been used in previous works as a proxy for the attractiveness of targets. The significance of their positive effect on residential burglary victimisation rate has been mixed and varied depending on the study region and the statistical method used (Bernasco & Luykx 2003, Bernasco & Nieuwbeerta 2005, Clare et al. 2009, Townsley et al. 2015, 2016). The finding that the effect of affluence was weak in some studies can be explained by the fact that most burglars do not live in affluent areas and hence are not in their awareness spaces, i.e. operating in an affluent neighbourhood is for them an unfamiliar terrain and the risk of being caught is higher (Evans 1989, Rengert & Wasilchick 2010). Other measures of affluence that have been used include *house ownership rates* (Bernasco & Luykx 2003).

With regard to non-residential burglary, the literature is more sparse. An analysis of non-residential burglary in Merseyside county in the UK by Bowers & Hirschfield (1999) shows that non-residential facilities have a higher risk of both victimisation and repeat victimisation. In particular, sport and educational facilities have a disproportionately higher risk of being targeted compared to other types of facilities. In the crime survey of business owners in the UK, the retail sector is the most vulnerable to burglaries (gov.uk 2017). For our application, we will use points of interest database from Ordnance Survey which include retail outlets, eating and drinking venues, accommodation units, sport and entertainment facilities, and health and education institutions (Ordnance Survey (GB) 2018).

3.2.2. Effort, convenience

Using the framework of crime pattern theory (Brantingham & Brantingham 1993) and routine activity theory (Cohen & Felson 1979), offenders will prefer locations that are part of their routine activities or are convenient to them, i.e. they are in their activity or awareness spaces. The studies performed using the data on detected residential burglaries unanimously agree that areas *close to the offender's home* are more likely to get targeted (Bernasco & Nieuwbeerta 2005, Townsley et al. 2015, Menting et al. 2019, Clare et al. 2009). In the study based on a survey of offenders, Menting et al. (2019) argue that other awareness spaces than their residence play a significant role in target selection. These include previous addresses, neighbourhoods of their family and friends, as well as places where they work and go about their recreation and leisure.

As confirmed by numerous studies, the spatial topology of the environment plays a significant role in the choice of a target. Brantingham & Brantingham (1975) have shown that houses in the interior of a block are less likely to get burgled. Similarly, Townsley et al. (2015), Bernasco & Nieuwbeerta (2005) showed that *single-family dwellings* are more vulnerable to burglaries than multi-family dwellings such as blocks of flats. Beavon et al. (1994) studied the effects of the street network and traffic flow on residential burglary and found that crime was higher in *more accessible* and *more frequented* areas. Similarly, Johnson & Bowers (2010) show that main street segments are more likely to become a burglary target. Clare et al. (2009), Bernasco et al. (2015) showed that the presence of connectors such as train stations increases the likelihood of being targeted, while the so-called barriers such as rivers or highways decrease it.

3.2.3. Risk, likelihood of completion

In the social disorganisation theory of crime (Shaw & McKay 1942, Sampson & Groves 1989), it is argued that social cohesion induces collective efficacy. The effect of collective efficacy on crime is twofold. First, strong social control deters those who are thinking of committing one. Second, it decreases the chance of a successful completion once an offender has chosen to do so. This theory focuses on the impact that social deprivation, economic deprivation, family disruption, ethnic heterogeneity, and residential turnover have on the crime rates within an area. Most offenders live in disadvantaged areas and often commit a crime in their awareness spaces (minimise effort). The attraction to 'prosperous targets' applies mostly to the local context (maximise gain). On the other hand, when a neighbourhood has high social cohesion (also known as 'collective efficacy'), there is mutual trust among neighbours and residents are more likely to intervene on behalf of the common good (Sampson et al. 1997).

In the context of residential burglary, *ethnic diversity* has been shown to be positively related to burglary rates (Sampson & Groves 1989, Bernasco & Nieuwbeerta 2005, Bernasco & Luykx 2003, Clare et al. 2009). *Residential turnover* is another measure of collective efficacy. Although Bernasco & Luykx (2003) document a positive relationship between residential turnover and the burglary rates, results in Bernasco & Nieuwbeerta (2005), Townsley et al. (2015) do not confirm that hypothesis. *Socio-economic variation* among residents has been shown to be positively related to general crime rates (e.g. Sampson et al. (1997), Johnson & Summers (2015)), but it was either not considered or shown insignificant in the studies on burglary we have reviewed. Other indicators of social disorganisation and their effect on general crime rates (not only burglary) are the high rate of single-parent households, one-person households as well as younger households Bernasco (2014), Sampson et al. (1997), Andresen (2010).

3.3. Covariates selection

Based on the criminological overview above and the availability of covariates, we form four model specifications, from very rich representations to sparse ones. Table 1 shows the covariates used in each of the specifications.

Variables that represent density, i.e. given by the count per cell, are log-transformed to improve the fit. For the same reason, mean household income and mean house price are in log form. Indicators of heterogeneity are computed using the index of variation introduced in Agresti & Agresti (1978). These include ethnic heterogeneity and occupation variation within an area. Both are indicators of the lack of social cohesion. Subsequently, all variables were standardised to have zero mean and standard deviation of one.

The first specification, *specification 1*, is the richest representation and includes variables that are a proxy for the same phenomenon. For example, both household income and house price are a measure of affluence. This choice is deliberate as we use a shrinkage prior for the regression coefficients to choose the most relevant variables.

The second specification, *specification 2*, removes covariates that are strongly correlated to others or lack strong evidence in the criminological literature. We remove *owner-occupied dwellings* for its strong correlations with the house dwellings and the fraction of houses that are detached or semi-detached. We remove *house dwellings* due to high correlation with (semi-)detached houses and stronger theoretical backing for the latter (e.g.Bernasco & Nieuwbeerta (2005)). We remove the *urbanisation level* because of little empirical evidence found in the literature. Naturally, it acts as a proxy for where buildings are, which is accounted for to a large extent by households and points of interest variables. We remove *single-parent households* due to a high correlation with social housing and unemployment rate, and the latter two being preferable indicators of social disorganisation.

In the third specification, *specification 3*, we exclude the following variables on top of those excluded in specification 2. *Median age*, as a proxy for collective efficacy, is removed due to weak evidence in previous studies and other measures of collective efficacy already present: ethnic and socio-economic heterogeneity. *One-person households* and *accommodation POIs* are removed because of weak empirical evidence from previous studies. *Mean household income* is removed due to insufficient evidence from previous studies and an already present and more preferable measure of affluence – house price. *Social housing* variable is removed because of weak empirical evidence and a high correlation with unemployment.

In the last specification, *specification 4*, we additionally remove *unemployment rate* due to weak empirical support from previous studies. This specification aggregates all POIs into a single variable (including accommodation POIs). This is to remove the strong correlations between them. As a single variable, it signifies the level of social activity: retail, education, entertainment,

Table 1. Models specifications that are used throughout the evaluation of the proposed model.

	1	2	3	4
log households (count per cell)	●	●	●	●
log retail POIs (count per cell)	●	●	●	
log eating/drinking POIs (count per cell)	●	●	●	
log edu/health POIs (count per cell)	●	●	●	
log accommodation POIs (count per cell)	●	●		
log sport/entertainment POIs (count per cell)	●	●	●	
log POIs (all categories count per cell)				●
houses (fraction of dwellings)	●			
(semi-)detached houses (fraction of dwellings)	●	●	●	●
social housing (fraction of dwellings)	●	●		
owner-occupied dwelling (fraction of dwellings)	●			
single-parent households (fraction of households)	●			
one-person households (fraction of households)	●	●		
unemployment rate	●	●	●	
ethnic heterogeneity measure (index of variation)	●	●	●	●
occupation variation measure (index of variation)	●	●	●	●
accessibility (estimated by Transport for London)	●	●	●	●
residential turnover (ratio of residents who moved in/out)	●	●	●	●
median age	●	●		
log mean household income	●	●		
log mean house price	●	●	●	●
urbanisation index (proportion of urban area)	●			

etc.

4. Results

After discussing the modelling choices and experimental settings, we compare SAM-GLM model to the log-Gaussian Cox process (LGCP), based on the out-of-sample generalisation and crime hotspot prediction. For LGCP, we use the standard formulation with a Matèrn covariance function (see section B for full details). Lastly, we interpret the results obtained using the proposed method and show the relevance for obtaining criminological insights.

4.1. Evaluation and interpretation

4.1.1. Out-of-sample performance

Firstly, we evaluate the performance of the proposed and competing models using the Poisson likelihood of one-period-ahead data given the model parameters obtained from training data. The likelihood denotes how likely the observed data are for given parameters. For a given sample from the posterior distribution of the model parameters, $\boldsymbol{\phi}^{(s)}$, the average pointwise *held-out log-likelihood* is defined as

$$\text{Held-out log likelihood} = \frac{1}{N}\sum_{n=1}^{N} \log p(\tilde{y}_n|\boldsymbol{\phi}^{(s)}), \tag{7}$$

where $p(\cdot)$ is the Poisson density function, $\tilde{y}_n$ is the realised next-period value. Log-likelihood is a relative measure used for model comparison and can only be used to compare models within the same family of models, in our case, Poisson-based models. A higher value indicates superior predictive power.

Next, we use the *root mean square error* (RMSE) metric. It is independent of the model and is measured at the same scale as the target variable. Given a sample from the posterior distribution of the model parameters, $\boldsymbol{\phi}^{(s)}$, we obtain a sample from the joint predictive probability distribution for the counts at all N locations, $\mathbf{y}^{(s)}$, using the sampling distribution of the data, $p(\mathbf{y}|\boldsymbol{\phi}^{(s)})$. Then, using the realised next-period value, $\tilde{\mathbf{y}} = (\tilde{\mathrm{y}}_1, \ldots, \tilde{\mathrm{y}}_N)$, the RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mathrm{y}_n^{(s)} - \tilde{\mathrm{y}}_n)^2}. \tag{8}$$

A lower value of RMSE indicates a better predictive performance.

4.1.2. Hotspot prediction

Given that burglary is our object of interest, we also evaluate models with respect to their ability to effectively model areas of high intensity, so-called *hotspots*. The predictive accuracy index (PAI) and predictive efficiency index (PEI) are two standard approaches in criminology for assessing the ability to predict crime hotspots.

PAI, introduced by Chainey et al. (2008), assesses the ability to capture as many crime instances as possible with the as little area as possible. For a given size of the area to be marked as hotspots, a, it is defined as

$$\mathrm{PAI} = \frac{c_a/C}{a/A},$$

where A is the total area of the study region, c_a is the number of crimes in the flagged hotspots with the total area a, and C is the total number of crimes in the study region.

However, for certain types of crime that are more serious and less frequent, it is important that each instance of crime is captured. PEI measures how effective the model forecasts are compared to what would a perfect model predict for a given size of the area to be marked as hotspots, a (Hunt 2016). It is defined as

$$\mathrm{PEI} = \frac{c_a}{c_a^*},$$

where c_a is the number of crimes in the hotspots of size a flagged by the model, and c_a^* is the maximum number of crimes that could have been captured using an area of size a.

In our context of a regular grid, we use both measures to compare competing models when up to n cells are flagged as hotspots. For a given n, a higher value indicates better hotspot prediction ability.

4.1.3. Interpretation of results

Estimating the effects of different spatial covariates helps us understand the underlying mechanisms of the point pattern.

In the mixtures of regressions literature, the interpretation of the individual regression coefficients is of no interest, or the focus is on reporting the regression coefficients ($\boldsymbol{\beta}_k$) for each component and quantifying their uncertainty so that their significance can be assessed (Frühwirth-Schnatter et al. 2019, ch. 8). To further interpret the coefficients, one could look at each mixture component specifically and interpret the coefficients in a classical way, conditional on the partitioning of observations. For example, for a GLM with the exponential link function, increasing a covariate by 1 unit multiplies the mean value of the observed variable by the exponential of the regression coefficient for that covariate, provided other covariates are held constant. However, this approach only allows component-specific conclusions as it depends on the distribution of the covariate for the associated component. For example, one mixture component may be active in areas with very small values for a specific covariate, while some other component is active in areas with high values. Comparing regression coefficients for that covariate across different components would not be appropriate.

Instead, to be able to compare the covariates across mixture components, we derive a covariate importance measure (`IMP`) that is motivated by the coefficient of determination, R^2. The objective of this measure is to assess the magnitude and the sign (positive/negative) of the effect of a covariate for a specific mixture component on the data fit. We measure the magnitude of the effect for a covariate j of the mixture component k as the ratio of the sum of squared residuals for the full model and the sum of squared residuals for the same model without covariate j, which is then subtracted from one. For a component k and a covariate j,

$$\texttt{IMP}_{kj} = 1 - \frac{\sum_n I(\mathrm{z}_n = k)(\mathrm{y}_n - \hat{\mathrm{y}}_{n\tilde{\boldsymbol{\beta}}})^2}{\sum_n I(\mathrm{z}_n = k)(\mathrm{y}_n - \hat{\mathrm{y}}_{n\bar{\boldsymbol{\beta}}^j})^2}, \tag{9}$$

where, $I(\mathrm{z}_n = k)$ is the indicator function of whether cell n is allocated to component k, $\hat{\mathrm{y}}_{n\tilde{\boldsymbol{\beta}}}$ is the predicted count using the full vector of inferred regression coefficients, and $\hat{\mathrm{y}}_{n\bar{\boldsymbol{\beta}}^j}$ is the predicted count using the regression coefficients with the jth coefficient set to zero. The magnitude of `IMP` is interpreted as a measure of the relative importance of the corresponding covariate for the model fit. A value of `IMP` closer to 1 represents that removing the corresponding covariate is more detrimental to model fit.

We determine the sign of `IMP` for a given covariate and a mixture component by inspecting the distribution of the covariate for the given component. We need to be careful with negative values as our covariates are centred around zero and standardised. To obtain the sign, we take the mean of the covariate across the cells that are allocated to the given component, and if that is positive, we take the sign of the corresponding β_{kj} estimate. Otherwise, we take the negative of the sign of the β_{kj} estimate.

4.2. Simulation study details

For the methodology developed in section 2, we need to choose the grid size, blocking structure, number of mixture components (K) and model specification.

4.2.1. Model choices

To choose grid size, we take into account the precision of the burglary point pattern. The published data have been anonymised by mapping exact locations to predefined (snap) points (police.uk 2018). We follow the recommendations in Tompson et al. (2015) who assess the accuracy of the anonymisation method by aggregating both the original and obfuscated data to areal counts at different resolutions and looking at the difference. They show that the aggregation at lower super output area (LSOA) level does not suffer from the bias introduced by the anonymisation process. Therefore, for our cell size, we approximately match an average-size LSOA to avoid the loss of precision caused by the anonymisation process. As a result, our grid has $N = 9824$ cells, each of which corresponds to an area of 400×400 metres.

For the blocking structure, we take advantage of the existing census output areas, that are designed to group homogeneous groups of households and people together (Office for National Statistics 2019). Given that our grid is approximately at the LSOA level, we choose MSOAs as the blocking structure. We assess the sensitivity of this choice in section 4.4.

The number of components, K, is a crucial parameter of our model. We run our model for varying K and use the performance measures introduced above to decide on the optimal number of components. From our experience, after a certain number of components, interpretation becomes harder while performance does not significantly improve.

We choose model specification based on the four options mentioned in section 3.3.

4.2.2. Dependence of blocks

In section 2 we have proposed two possible formulations for the prior on the mixture weights: the multinomial logit transformation of K Gaussian random fields and independent Dirichlet random variables. To assess whether assuming block dependence has a major effect on the quality of the model, we compare the out-of-sample performance for both variants of the model. For this comparison, we set the blocking scheme to MSOA, use model specification 4, and estimate the

Table 2. Model performance comparison of two variants of the model – dependent blocks using the logistic transform of K Gaussian processes, and independent blocks with Dirichlet prior. Reported values are a mean and standard deviation obtained from MCMC samples. Blocking: MSOA, training data: burglary 2015, test data: 2016, model specification 4.

K	Held-out log-likelihood		RMSE	
	Independent	Dependent	Independent	Dependent
2	-2.607 ± 0.010	$\mathbf{-2.605 \pm 0.010}$*****	$\mathbf{4.999 \pm 0.028}$*****	5.010 ± 0.028
3	-2.598 ± 0.012	$\mathbf{-2.593 \pm 0.011}$*****	4.973 ± 0.036	$\mathbf{4.950 \pm 0.031}$*****
4	$\mathbf{-2.588 \pm 0.011}$*****	-2.606 ± 0.012	$\mathbf{4.964 \pm 0.034}$*****	4.988 ± 0.031

model on the burglary 2015 dataset. To fit the model with dependent blocks, we use the squared exponential kernel (Rasmussen & Williams 2006) where we choose the lengthscale parameter by optimising out-of-sample RMSE using grid search. Table 2 shows the mean and the standard deviation of the samples of held-out log-likelihood and RMSE for both variants of the model, and for different values of K. The bold typeface signifies which method performed better for the given K and for the given metric. The star indicates statistical significance with p-value $< 10^{-3}$ obtained from a two-sample t-test of samples of each metric for each variant of the model.

The results in table 2 show that the model with dependent blocks does not consistently lead to improved performance. This indicates that block dependence structure in the burglary point pattern data that we consider is not a major effect. These findings highlight some aspects of the data structure in terms of capturing these effects and suggest that the point pattern data at a higher precision would be needed to uncover these effects, if they are present. For this reason, in the rest of the paper we only consider independent blocks with Dirichlet prior weights as described in section 2.3.

4.2.3. Identifiability

As mentioned in section 2, the traceplot of the log-likelihood can be inspected for label-switching. From our experience, the sampler would choose one of the $K!$ modes, that are a consequence of the likelihood invariance, and is unlikely to switch to another mode due to the high dimensionality of the parameter space.

4.3. SAM-GLM performance

Figures 2 and 3 report performance for the 2015 and 2013-2015 datasets, respectively. On the left panels of the figures, we report the box-plot of the posterior distribution of the average held-out log-likelihood. We show the box-plot for different model specifications for both SAM-GLM with an increasing number of components (K) and LGCP models. On the right panels, we report analogous plots for the root mean square error metric (RMSE).

For the one-year dataset, SAM-GLM model matches the predictive performance of the LGCP model for $K = 2$ components on both metrics. For the three-year dataset, $K = 3$ components are enough to match the LGCP model using the held-out log-likelihood, but at least $K = 4$ components are required for RMSE. The extra components required to match the performance of LGCP could be explained by the fact that the three-year point pattern will naturally be smoother and thus easier to interpolate non-parametrically using the Gaussian random field part of LGCP. The probability distribution for both metrics and for all models are more concentrated for the three-year dataset. For the one-year dataset, it is clear that $K = 2$ or $K = 3$ is the optimal number of components. For the three-year counterpart, the range between 3 and 5 components would be an appropriate choice. For both datasets, the performance does not vary significantly for different model specifications. Consequently, in the following sections, we limit our attention to specification 4 due to its parsimony.

While out-of-sample performance, measured by the held-out log-likelihood or RMSE, takes into account all locations, practitioners might only be interested in predicting crime hotspots. To

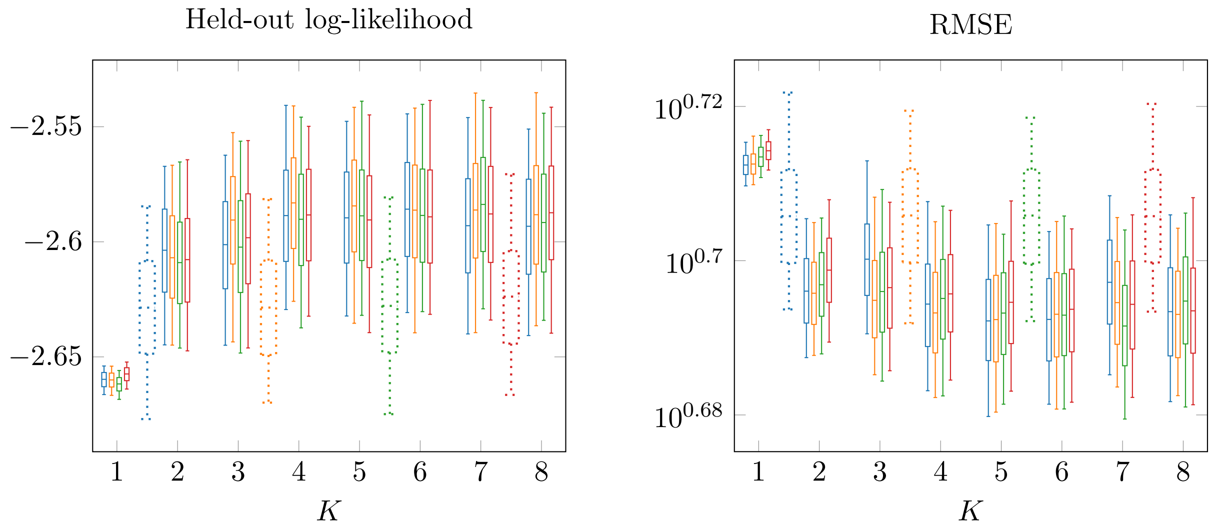


Figure 2. Evaluation of the performance of SAM-GLM (—), compared to LGCP (·····) for the one-year dataset. Log-likelihood and root mean square error for the held-out data are shown for different model specifications: specification 1 (—), specification 2 (—), specification 3 (—), specification 4 (—). Blocking: MSOA, training data: burglary 2015, test data: burglary 2016. Note that the axis with the value of K does not apply to the LGCP results.

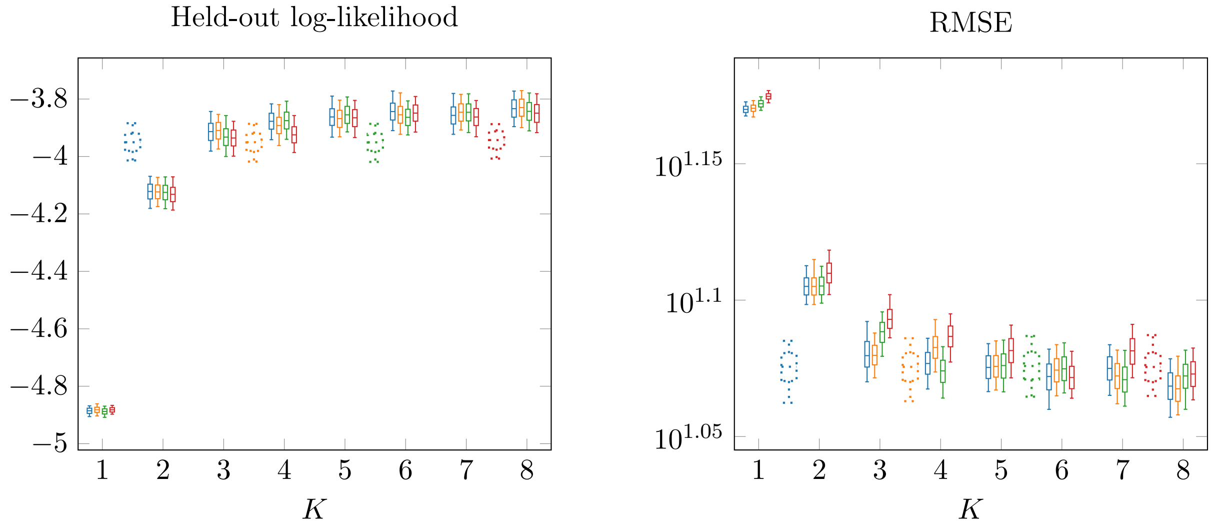


Figure 3. Evaluation of the performance of SAM-GLM (—), compared to LGCP (·····) for the three-year dataset. Log-likelihood and root mean square error for the held-out data are shown for different model specifications: specification 1 (—), specification 2 (—), specification 3 (—), specification 4 (—). Blocking: MSOA, training data: burglary 2013-2015, test data: burglary 2016-2018. Note that the axis with the value of K does not apply to the LGCP results.

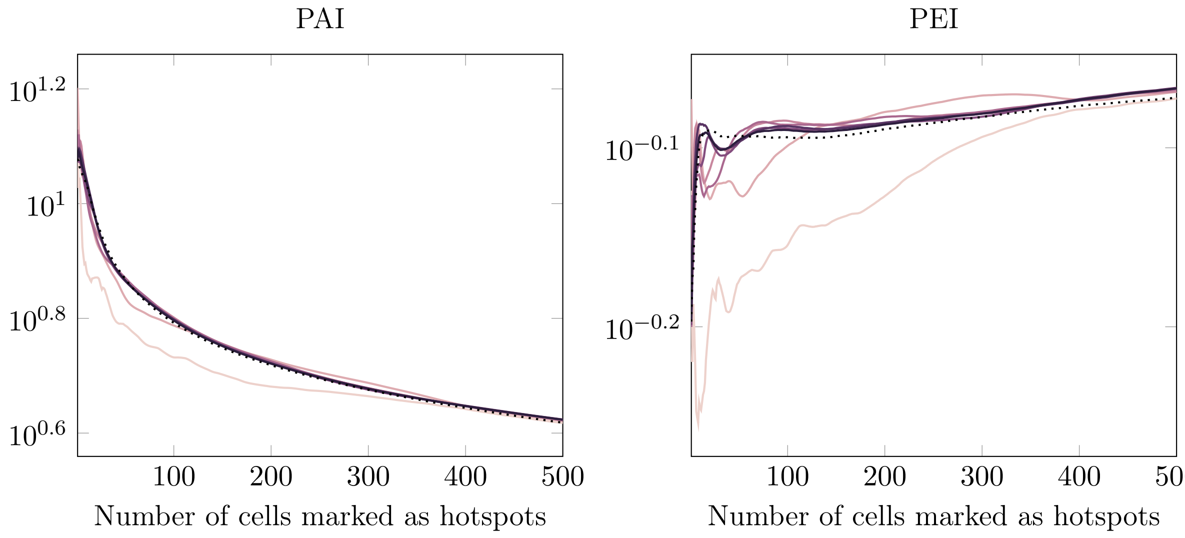


Figure 4. PAI/PEI performance SAM-GLM (—) and LGCP (·····) models, using specification 4. For the SAM-GLM results, the colour of the line represents the number of components: $K = 1$(—), $K = 2$(—), $K = 3$(—), $K = 4$(—), $K = 5$(—), $K = 6$(—), $K = 7$ (—). Blocking: MSOA, training data: burglary 2015, test data: burglary 2016, model specification: 4.

this end, we evaluate PAI and PEI (see section 4.1) as measures of hotspot prediction. Figures 4 and 5 show the plots of PAI and PEI measures for both models with specification 4, using the 2015 and 2013-2015 datasets, respectively. The plots show the score for when up to 500 cells (around 5% of the study region) are flagged as hotspots. Hotspots are chosen as the n cells with the highest expected value of burglaries. For the one-year dataset, the SAM-GLM model with $K = 2$ components is enough to outperform LGCP on PEI measure when between 50 and 500 cells are flagged as hotspots. For PAI measure, no significant difference can be seen for $K > 2$. The results based on the three-year data favour LGCP model when up to 150 cells are flagged as hotspots and $K < 5$. After adding more components, the SAM-GLM performance matches that of LGCP. When between 150 and 500 cells are flagged, $K \geq 3$ components is enough to outperform LGCP. These results are consistent with the previous finding that outperforming LGCP on the three-year dataset requires more components.

4.4. Block size sensitivity

The proposed model requires a specification of the blocking structure for the mixture weights prior. To assess sensitivity of this choice, we compare to local authority districts (LAD), as well as a single block for the whole study region. In the latter case, the model reduces to a non-spatial mixture of Poisson GLMs. There are 946 MSOAs, and 33 LADs in the study region. The structure is hierarchical – multiple non-overlapping contiguous MSOAs constitute single LAD region.

Figures 6, and 7 show the box-plots of the held-out log-likelihood and RMSE for the one-year and the three-year datasets, respectively. The results for both metrics indicate that imposing spatial information using more localised prior results in better out-of-sample performance for the one-year dataset. To confirm that the difference is statistically significant, we performed an unpaired two-sample t-test comparing RMSE samples obtained using MSOA blocking structure to those obtained using the LAD and single blocks, respectively. Table 3 summarises the t-statistics and p-values. For the three-year dataset, there is no evident difference, and spatial prior does not improve predictive performance of the model. This is not surprising as the 3-year observation window will provide more information and thus the model is less likely to overfit even if we do not impose spatial dependence within the blocks.

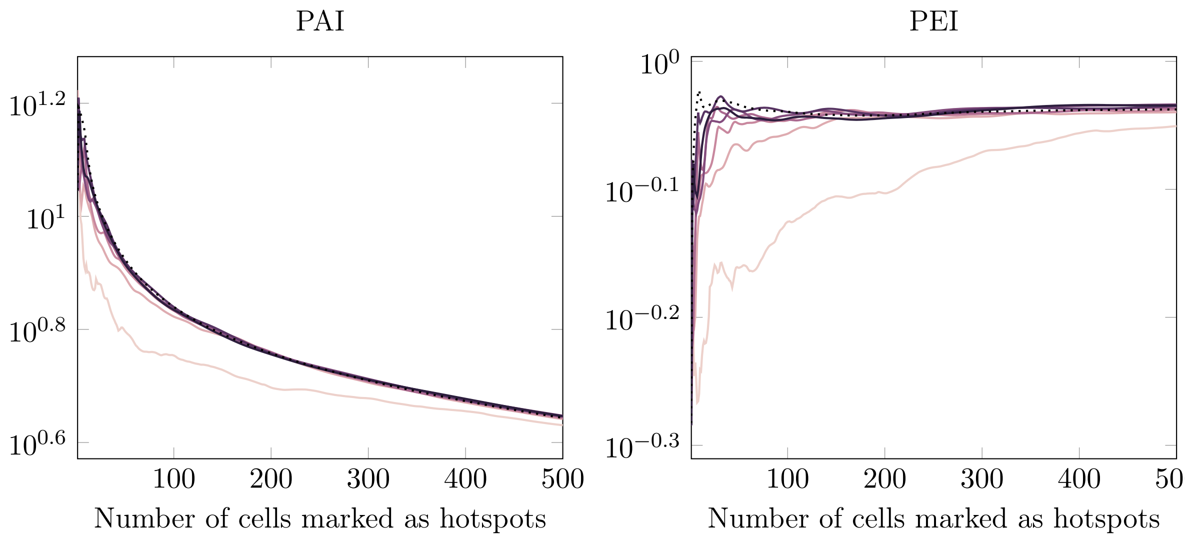


Figure 5. PAI/PEI performance SAM-GLM (—) and LGCP (·····) models, using specification 4. For the SAM-GLM results, the colour of the line represents the number of components: $K = 1$(—), $K = 2$(—), $K = 3$(—), $K = 4$(—), $K = 5$(—), $K = 6$(—), $K = 7$ (—). Blocking: MSOA, training data: burglary 2013-2015, test data: burglary 2016-2018, model specification: 4.

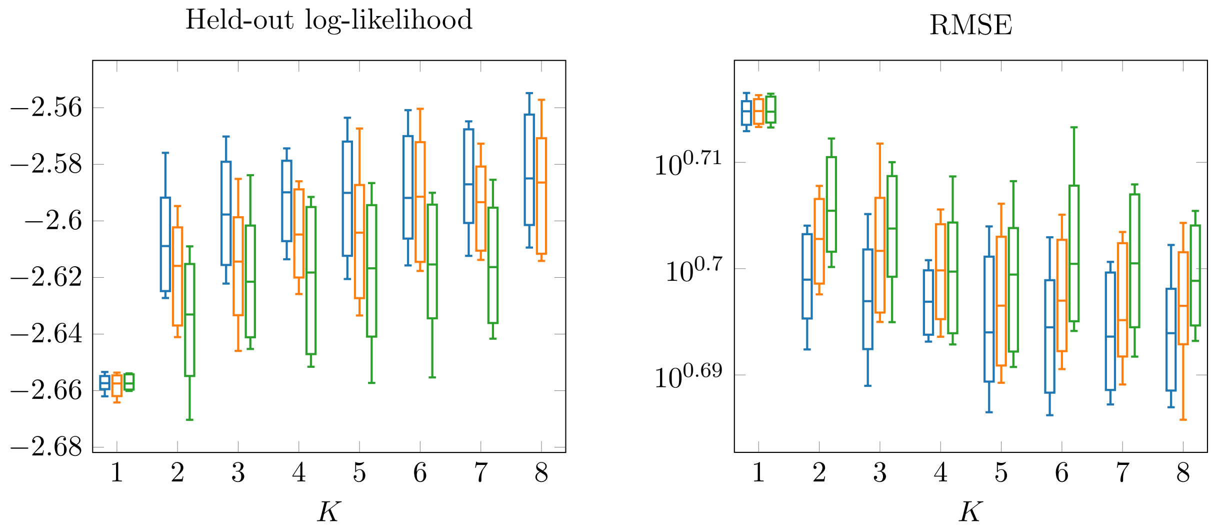


Figure 6. Log-likelihood and root mean square error for the held-out data for different block sizes: MSOA(—), LAD(—), single block(—). The error bars represent the standard deviation obtained from the respective MCMC samples. Training data: 2015, test data: 2016, model specification 4

Table 3. Sensitivity analysis of block sizes. p-values comparing whether the difference in RMSE performance is significant. Training data: burglary 2015, test data: burglary 2016, specification 4.

K	MSOA vs LAD		MSOA vs SINGLE	
	t-statistic	p-value	t-statistic	p-value
2	-68.732	$< 10^{-3}$	-115.042	$< 10^{-3}$
3	-76.260	$< 10^{-3}$	-87.534	$< 10^{-3}$
4	-39.016	$< 10^{-3}$	-35.207	$< 10^{-3}$
5	-26.858	$< 10^{-3}$	-52.991	$< 10^{-3}$
6	-41.913	$< 10^{-3}$	-76.152	$< 10^{-3}$
7	-12.173	$< 10^{-3}$	-56.847	$< 10^{-3}$
8	-31.547	$< 10^{-3}$	-66.688	$< 10^{-3}$

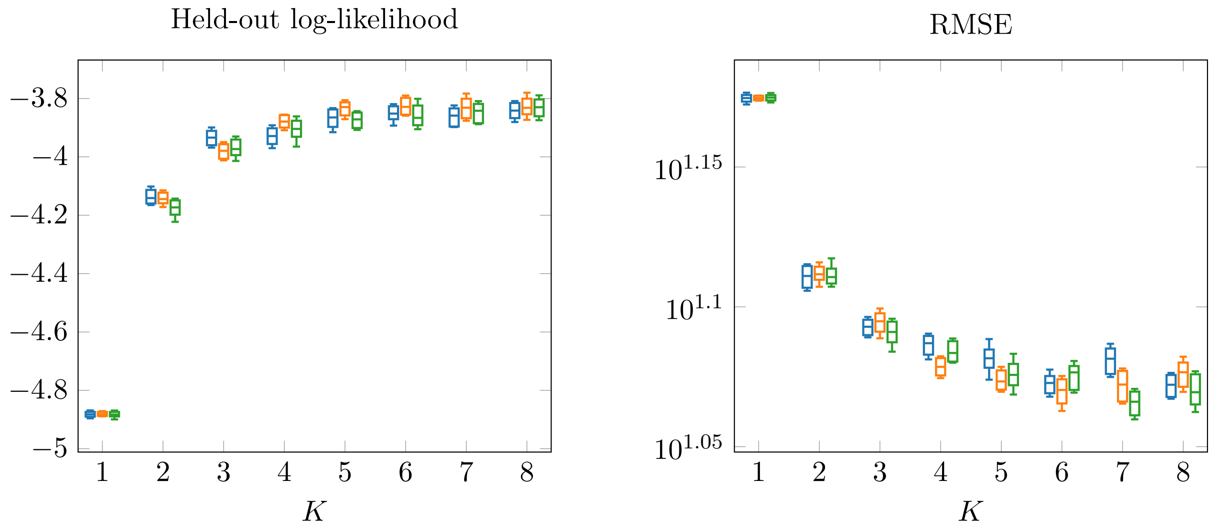


Figure 7. Log-likelihood and root mean square error for the held-out data for different block sizes: MSOA(—), LAD(—), single block(—). The error bars represent the standard deviation obtained from the respective MCMC samples. Training data: 2013-2015, test data: 2016-2018, model specification 4

4.5. *Interpretation*

For this analysis, we choose the three-year dataset because more data will lead to more robust inferences of the parameters. We choose specification 4 with $K = 3$ components because of its parsimony and the excellent performance shown above – for the three-year dataset and specification 4, there does not seem to be a significant improvement after $K > 3$ components. Figure 8 shows the component allocation maps and the `IMP` measure with the effect sign $(+/-)$ for each covariate for all the three components. The allocation map for each component shows the proportion of the MCMC samples a cell is allocated to that component. The alphanumeric labels on the allocation plots are used in the discussion below when referring to specific locations. `IMP` is computed for each sample and component separately and then averaged over the MCMC samples. We also report the standard deviation of the `IMP` estimate in brackets.

The first component is active throughout the study region, with large clusters around residential areas. These include areas around Kensington, Fulham, and Shepherd's Bush (A); Hounslow, Kingston, Richmond, and Twickenham (2); Hayes and Southall (C); Harrow and Edgware (D); East Barnet, Enfield, Walthamstow, Wood Green (E); Barking and Dagenham (F); Bexley (G); Orpington (H); Bromley (I); Croydon, and Purley (J); New Malden, and Morden (K). In this component, the number of households and points of interest have the strongest effect (excluding the intercept) – burglaries happen where targets are. Accessibility has also been inferred as an important covariate, consistent with the past criminological studies. In this component, house price is inferred as having a positive effect on the intensity of burglary, suggesting that offenders choose attractive targets. The positive effect of ethnic heterogeneity confirms the hypothesis from the social disorganisation theory. The other indicators of social disorganisation – occupation variation, residential turnover – are weaker but are consistent with the existing criminology literature. House price as a measure of reward and the proportion of houses that are detached and semi-detached have low `IMP` value.

Component 2 is active in the city centre and in the high streets of neighbourhoods: Soho, Mayfair, Covent Garden, Marylebone, Fitzrovia (L); Shoreditch and Stratford (M); Streatham and Tooting Bec (N); Wembley, and Brent (O); Enfield, Hampstead (P); Romford (Q); Orpington (R); Wembley, Harrow (S). Burglary rates in these locations are largely driven by points of interest and households. Compared to the first component (residential), the magnitudes of `IMP` values for these covariates are different - points of interest are more important for this component, and the number of households is more important for the first component. Accessibility measure is inferred to have high importance in this component. This measure is high in the city centre and around the high streets, which are usually well-connected to the public transport system. This confirms findings from crime pattern theory and routine activity theory which suggest that offenders choose locations that are part of their usual routine and in their awareness spaces. Ethnic heterogeneity and occupation variation have strong positive effect and signify the lack of social cohesion. Unexpectedly, our model infers a negative relationship between residential turnover and burglary intensity. Association of high residential turnover with the reduced risk of burglary apprehension has been shown as significant in only a few studies and was limited to *residential* burglary (Bernasco & Luykx 2003, Bernasco & Nieuwbeerta 2005, Townsley et al. 2015). Areas that are less residential such as high streets have a higher proportion of flats. Dwellings with shared premises such as flats have been shown to less likely become a target than one-household buildings (Beavon et al. 1994). Another possible reason could be the staleness of the data for the covariates which are taken from the 2011 census. Also, house price has been inferred to have a negative effect, i.e. more affluent locations are less likely to get targeted. This is contrary to the first component. A possible explanation mentioned in previous studies is that offenders often live in disadvantaged areas and choose targets within their awareness spaces, which are less likely to be affluent areas (Evans 1989, Rengert & Wasilchick 2010).

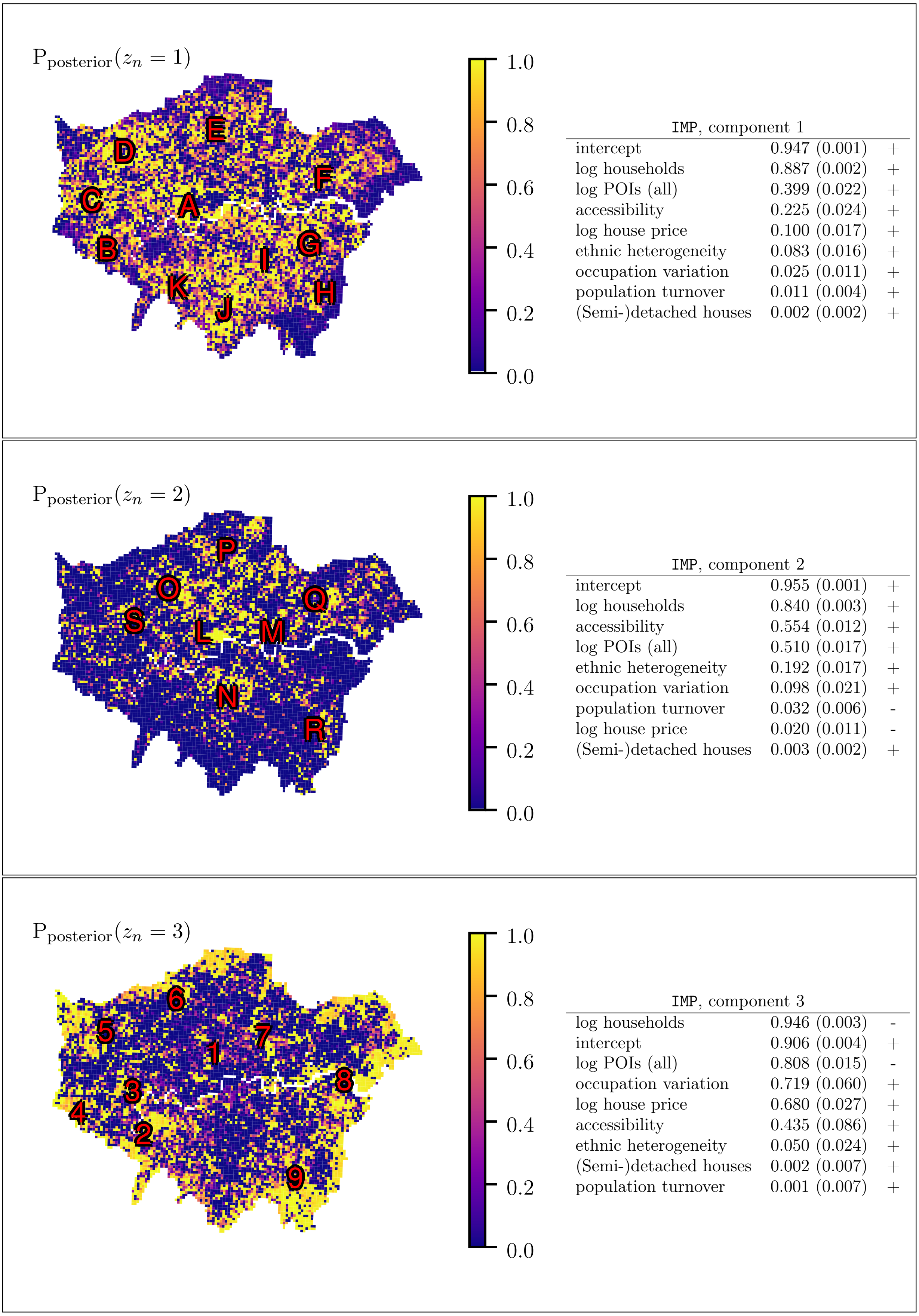
The last component is active in the areas of low intensity. These include Hyde Park, Regent's Park, Hampstead Heath (1); Richmond and Bushy parks (2); Osterley Park and Kew botanic gardens (3); Heathrow airport (4); RAF Northolt, and parks near Harrow (5); Edgware fields (6); Lee Valley (7); industrial zone in Barking and Rainham Marshes (8); parks around Bromley and Biggin Hill airport (9); and other non-urban areas located on the edges of the map. This component explains locations with little criminal activity, signified by negative `IMP` for the number of households and points of interest. Occupation variation, as a measure of socioeconomic heterogeneity, is strongly positive, which would support the hypothesis from social disorganisation theory. However, this is more likely due to the very low population in those areas which results in high occupation variation measure. Accessibility measure also has a positive effect on burglary rates in these locations. This is expected and in line with the hypotheses from the crime pattern theory. Other covariates have very small `IMP` values.

The allocation of cells partitions the map into three clusters. By aggregating the number of observed crimes that occurred in each cluster we get that components 1, 2, and 3 cover 46%, 42%, 12% of all burglaries during the 2013-2015 period, respectively. Official aggregated police data for this period make the split of 64% and 36% for residential and non-residential burglary (police.uk 2019). Our inference agrees that there is more residential burglary than non-residential burglary and that approximately 35-45% of burglaries are non-residential. It is unclear whether the crime in low-count areas, which according to our model accounts for 12%, is residential or non-residential.

The support for the presence of spatial heterogeneity is further given by inspecting the inferences made by the LGCP model (for LGCP details see section B in the appendix). The left panel of figure 9 shows standard deviations of the marginal posterior distributions of the Gaussian random field component (f). It is clear that the variance of the field component is clustered, where the regions with higher values are easily identifiable as those less urbanised. In contrast, SAM-GLM model has pickled up this heterogeneity by allowing a separate component for it (see component 3 in figure 8). The right panel of figure 9 shows `IMP` computed for all components of the LGCP model. `IMP` measure for the field component of the model is computed by treating it as a covariate with the coefficient equal to one. The `IMP` value for the latent field component is the third-highest, after the intercept and the number of households. A large contribution from the latent component indicates that the linear term in the Poisson regression model cannot on its own sufficiently explain the variation in the intensity of burglary.

4.6. Overdispersion, excess of zeros

The discussion of the inferences above shows that our model effectively handles excess of zeros by allocating low-count cells (non-urban areas) its own cluster, which has its own regression coefficients. Similarly, the proposed mixture model is able to reduce the overdispersion problem



$\mathrm{P}_{\mathrm{posterior}}(z_n = 1)$

`IMP`, component 1		
intercept	0.947 (0.001)	+
log households	0.887 (0.002)	+
log POIs (all)	0.399 (0.022)	+
accessibility	0.225 (0.024)	+
log house price	0.100 (0.017)	+
ethnic heterogeneity	0.083 (0.016)	+
occupation variation	0.025 (0.011)	+
population turnover	0.011 (0.004)	+
(Semi-)detached houses	0.002 (0.002)	+

$\mathrm{P}_{\mathrm{posterior}}(z_n = 2)$

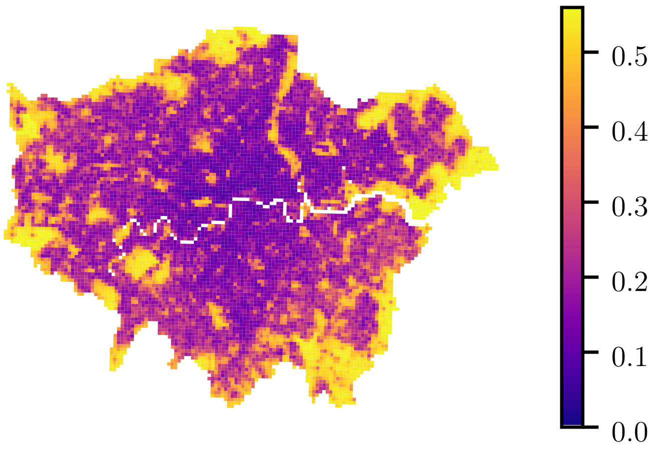
`IMP`, component 2		
intercept	0.955 (0.001)	+
log households	0.840 (0.003)	+
accessibility	0.554 (0.012)	+
log POIs (all)	0.510 (0.017)	+
ethnic heterogeneity	0.192 (0.017)	+
occupation variation	0.098 (0.021)	+
population turnover	0.032 (0.006)	-
log house price	0.020 (0.011)	-
(Semi-)detached houses	0.003 (0.002)	+

$\mathrm{P}_{\mathrm{posterior}}(z_n = 3)$

`IMP`, component 3		
log households	0.946 (0.003)	-
intercept	0.906 (0.004)	+
log POIs (all)	0.808 (0.015)	-
occupation variation	0.719 (0.060)	+
log house price	0.680 (0.027)	+
accessibility	0.435 (0.086)	+
ethnic heterogeneity	0.050 (0.024)	+
(Semi-)detached houses	0.002 (0.007)	+
population turnover	0.001 (0.007)	+

Figure 8. Mixture model, allocations and `IMP` table for each mixture component. Training data: 2013-2015, specification 4.

Standard deviation of the posterior of $\mathbf{f}$ (LGCP)



	IMP, LGCP	
intercept	0.953 (0.001)	+
log households	0.937 (0.002)	+
field	0.810 (0.005)	+
log POIs (all)	0.728 (0.013)	+
accessibility	0.103 (0.058)	+
occupation variation	0.097 (0.052)	+
log house price	0.089 (0.043)	+
ethnic heterogeneity	0.061 (0.029)	+
population turnover	0.002 (0.027)	+
(Semi-)detached houses	0.000 (0.027)	-

Figure 9. Left: Standard deviation of the posterior distribution of the latent field, $\mathbf{f}$, of the LGCP model. It is clear that, it is clustered and the elevated levels correspond to non-urban locations, airports, and parks (see the discussion above). Right: `IMP` measure for the component of the LGCP model. For both panels, training data: 2013-2015, model specification: 4.

that is present in the standard Poisson GLM model (the special case of SAM-GLM, with $K = 1$). The mixture model may allocate each cell to a cluster that better describes the burglary count in that location. Inspecting the Pearson χ^2 statistic ($\chi^2 = \sum_{i=1}^{N} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$) provides supporting evidence for this. Introduction of two extra components has resulted in the 81% decrease, from 106 942.43 to 20 028.99, showing a better model fit. This is further confirmed by a scatter plot of expected vs observed counts for the Poisson GLM model and the proposed model with $K = 3$ as shown in figure 11 in the appendix.

5. Conclusions

Spatial point patterns on large spatial regions, such as metropolitan areas, often exhibit localised behaviour. Motivated by this, we propose a mixture model that accounts for spatial heterogeneity as well as incorporates spatial dependence. Each component of the mixture is a model in itself, and thus allows for different locations to follow a different model, e.g. in the urban context, less-urbanised locations can assume a different model from the city centre. Each component is an instance of the generalised linear model (GLM) which includes covariates. We account for spatial dependence through the mixture allocation part. The allocation of each location to one of the components is informed by both the data and the prior information. By utilising existing blocks structure, or defining a custom one, the prior supports locations within the same block to come from the same component. This formulation attempts to find the right balance between the ability to model sharp spatial variations and borrowing statistical strength for locations within the same block. Additionally, the model allows for spatial dependence between the blocks. Following the Bayesian framework, we present a Markov Chain Monte Carlo sampler to infer the posterior distributions. Inspection of the posterior distributions of the model parameters allows us to learn new insights about the underlying mechanisms of the point pattern.

Our results show that London burglary data are effectively modelled by the proposed method. Using out-of-sample and crime hotspot prediction evaluation measures, we show our model outperforms log-Gaussian Cox process (with Matèrn covariance function) that is the default model for point processes and is more computationally tractable.

The focus of this work on burglary crime does not limit the potential uses of the proposed model. We believe that the model can be applied in a wider setting of analysing spatial point patterns that may show localised behaviour and heterogeneity.

Future analysis could consider several directions not explored in this work. Firstly, our inference scheme for the model with block dependence produces an $\mathcal{O}\big(B^3 \times K\big)$ algorithm. To reduce this complexity, one could consider K level sets of a single Gaussian random field for mixture weights, instead of K Gaussian fields, thus reducing dimensionality (Hildeman et al. 2018, Fernández & Green 2002). Another approach is assuming Markovian structure of the Gaussian random fields, resulting in sparse computational methods(Rue & Held 2005). A different approach is considering inference schemes that are less computationally demanding than MCMC such as variational methods (Jordan et al. 1999). Secondly, different options for specifying the term that involves covariates could be explored. One could consider forcing certain covariates to share the coefficients across all components if there is a strong prior belief for doing so. Another possible area of investigation is spatially varying coefficient processes method, proposed by Gelfand et al. (2003).

6. Acknowledgements

We thank the Associate Editor and the anonymous referees for their helpful comments and suggestions. All authors were supported by the EPSRC grants EP/P020720/1 and EP/P020720/2 for Inference, COmputation and Numerics for Insights into Cities (ICONIC, `https://iconicmath.org/`). Jan Povala was also supported by EPSRC (EP/L015129/1). Mark Girolami was supported by EPSRC grants EP/T000414/1, EP/R018413/2, EP/R034710/1, EP/R004889/1, and a Royal Academy of Engineering Research Chair in Data Centric Engineering. Parts of this work were carried out while Jan Povala was visiting Strategic Insights Unit at the Metropolitan Police Service in London. The authors are also grateful to Louis Ellam for insightful comments and suggestions, and to Pavol Povala for his feedback on the drafts of the manuscript.

7. Implementation

The source code that implements the methodology and reproduces the experiments is available at `https://github.com/jp2011/spatial-poisson-mixtures`.

References

Abramowitz, M. & Stegun, I. A. (1965), *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, Vol. 55, Courier Corporation.

Agresti, A. & Agresti, B. F. (1978), ‘Statistical Analysis of Qualitative Variation’, *Sociological Methodology* **9**, 204.

Aldor-Noiman, S., Brown, L. D., Fox, E. B. & Stine, R. A. (2017), ‘Spatio-temporal low count processes with application to violent crime events’, *Statistica Sinica* .

Alvares, D., Armero, C. & Forte, A. (2018), ‘What does objective mean in a dirichlet-multinomial process?’, *International Statistical Review* **86**(1), 106–118.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12231*

Andresen, M. A. (2010), The place of environmental criminology within criminological thought, *in* ‘Classics in environmental criminology’, CRC Press, pp. 21–44.

Anselin, L., Cohen, J., Cook, D., Gorr, W. & Tita, G. (2000), ‘Spatial analyses of crime’, *Criminal justice* **4**(2), 213–262.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2015), *Hierarchical modeling and analysis for spatial data*, number 135 *in* ‘Monographs on statistics and applied probability’, second edition edn, CRC Press, Taylor & Francis Group, Boca Raton.

Beavon, D. J., Brantingham, P. L. & Brantingham, P. J. (1994), ‘The influence of street networks on the patterning of property offenses’, *Crime prevention studies* **2**, 115–148.

Bernasco, W. (2014), Residential Burglary, *in* G. Bruinsma & D. Weisburd, eds, ‘Encyclopedia of Criminology and Criminal Justice’, Springer New York, New York, NY, pp. 4381–4391.

Bernasco, W., Johnson, S. D. & Ruiter, S. (2015), ‘Learning where to offend: Effects of past on future burglary locations’, *Applied Geography* **60**, 120–129.

Bernasco, W. & Luykx, F. (2003), 'Effects of Attractiveness, Opportunity, and Accessibility to Burglars on Residential Burglary Rates of Urban Neighbourhoods', *Criminology* **41**(3), 981–1002.

Bernasco, W. & Nieuwbeerta, P. (2005), 'How Do Residential Burglars Select Target Areas?', *The British Journal of Criminology* **45**(3), 296–315.

Bowers, K. & Hirschfield, A. (1999), 'Exploring links between crime and disadvantage in north-west England: an analysis using geographical information systems', *International Journal of Geographical Information Science* **13**(2), 159–184.

Brantingham, P. & Brantingham, P. (1981), Notes on the geometry of crime, *in* 'Environmental Criminology', Sage Publications, Beverly Hills, CA.

Brantingham, P. J. & Brantingham, P. L. (1975), 'The spatial patterning of burglary', *The Howard Journal of Criminal Justice* **14**(2), 11–23.

Brantingham, P. L. & Brantingham, P. J. (1993), 'Nodes, paths and edges: Considerations on the complexity of crime and the physical environment', *Journal of Environmental Psychology* **13**(1), 3–28.

Breslow, N. E. (1984), 'Extra-poisson variation in log-linear models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **33**(1), 38–44.

Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. (1996), 'Geographically weighted regression: a method for exploring spatial nonstationarity', *Geographical analysis* **28**(4), 281–298.

Cameron, A. & Trivedi, P. K. (1990), 'Regression-based tests for overdispersion in the Poisson model', *Journal of Econometrics* **46**(3), 347–364.

Celeux, G., Hurn, M. & Robert, C. P. (2000), 'Computational and Inferential Difficulties with Mixture Posterior Distributions', *Journal of the American Statistical Association* **95**(451), 957–970.

Chainey, S., Tompson, L. & Uhlig, S. (2008), 'The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime', *Security Journal* **21**(1-2), 4–28.

Clare, J., Fernandez, J. & Morgan, F. (2009), 'Formal Evaluation of the Impact of Barriers and Connectors on Residential Burglars' Macro-Level Offending Location Choices', *Australian & New Zealand Journal of Criminology* **42**(2), 139–158.

Clarke, R. V. & Cornish, D. B. (1985), 'Modeling Offenders' Decisions: A Framework for Research and Policy', *Crime and Justice* **6**, 147–185.

Cohen, L. E. & Felson, M. (1979), 'Social Change and Crime Rate Trends: A Routine Activity Approach', *American Sociological Review* **44**, 588–608.

Diggle, P. J., Moraga, P., Rowlingson, B. & Taylor, B. M. (2013), 'Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm', *Statistical Science* **28**(4), 542–563.

Duane, S., Kennedy, A., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid monte carlo', *Physics Letters B* **195**(2), 216 – 222.

Evans, D. J. (1989), Geographical Analyses of Residential Burglary, *in* D. J. Evans & D. T. Herbert, eds, 'The Geography of Crime', Routledge, London, pp. 86–107.

Felson, M. & Clarke, R. V. (1998), 'Opportunity makes the thief', *Police research series, paper* **98**, 1–36.

Fernández, C. & Green, P. J. (2002), 'Modelling spatially correlated data via mixtures: a Bayesian approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 805–826.

Flaxman, S., Chirico, M., Pereira, P. & Loeffler, C. (2019), 'Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-Time Crime Forecasting Challenge"', *Ann. Appl. Stat.* **13**(4), 2564–2585.

Flaxman, S., Wilson, A. G., Neil, D. B., Nickisch, H. & Smola, A. J. (2015), Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods, *in* 'Proceedings of the 32nd International Conference on International Conference on Machine Learning', Vol. 37 of *ICML'15*, JMLR.org, Lille, France, pp. 607–616.

Frühwirth-Schnatter, S., Celeux, G. & Robert, C. P., eds (2019), *Handbook of mixture analysis*, CRC Press, Boca Raton, Florida.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F. & Banerjee, S. (2003), 'Spatial Modeling With Spatially Varying Coefficient Processes', *Journal of the American Statistical Association* **98**(462), 387–396.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6), 721–741.

Girolami, M. & Calderhead, B. (2011), 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods: Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

gov.uk (2017), 'Crime against businesses: findings from the 2017 Commercial Victimisation Survey'.
URL: *https://www.gov.uk/government/statistics/crime-against-businesses-findings-from-the-2017-commercial-victimisation-survey*

Green, P. J. (2010), Introduction to Finite Mixtures, *in* S. Frühwirth-Schnatter, G. Celeux & C. P. Robert, eds, 'Handbook of Spatial Statistics', Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, Florida.

Green, P. J. & Richardson, S. (2002), 'Hidden Markov Models and Disease Mapping', *Journal of the American Statistical Association* **97**(460), 1055–1070.

Grün, B. & Leisch, F. (2008), Finite Mixtures of Generalized Linear Regression Models, *in* 'Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg', Physica-Verlag HD, Heidelberg, pp. 205–230.

Hildeman, A., Bolin, D., Wallin, J. & Illian, J. B. (2018), 'Level set Cox processes', *Spatial Statistics* **28**, 169–193.

Hunt, J. M. (2016), Do crime hot spots move? Exploring the effects of the modifiable areal unit problem and modifiable temporal unit problem on crime hot spot stability, PhD Thesis, American University, Washington, D.C.

Johnson, S. D. & Bowers, K. J. (2004), 'The Stability of Space-Time Clusters of Burglary', *The British Journal of Criminology* **44**(1), 55–65.

Johnson, S. D. & Bowers, K. J. (2010), 'Permeability and Burglary Risk: Are Cul-de-Sacs Safer?', *Journal of Quantitative Criminology* **26**(1), 89–111.

Johnson, S. D. & Summers, L. (2015), 'Testing Ecological Theories of Offender Spatial Decision Making Using a Discrete Choice Model', *Crime & Delinquency* **61**(3), 454–480.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999), 'An Introduction to Variational Methods for Graphical Models', *Machine Learning* **37**(2), 183–233.

Kleiber, C. & Zeileis, A. (2008), *Applied econometrics with R*, Springer-Verlag, New York. ISBN 978-0-387-77316-2.
URL: *https://CRAN.R-project.org/package=AER*

Knorr-Held, L. & Raßer, G. (2000), 'Bayesian Detection of Clusters and Discontinuities in Disease Maps', *Biometrics* **56**(1), 13–21.

McCullagh, P. & Nelder, J. A. (1998), *Generalized linear models*, number 37 *in* 'Monographs on statistics and applied probability', 2nd ed edn, Chapman & Hall/CRC, Boca Raton.

Menting, B., Lammers, M., Ruiter, S. & Bernasco, W. (2019), 'The Influence of Activity Space and Visiting Frequency on Crime Location Choice: Findings from an Online Self-Report Survey', *The British Journal of Criminology* p. In press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of State Calculations by Fast Computing Machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. (2011), 'Self-Exciting Point Process Modeling of Crime', *Journal of the American Statistical Association* **106**(493), 100–108.

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998), 'Log Gaussian Cox Processes', *Scandinavian Journal of Statistics* **25**(3), 451–482.

Møller, J. & Waagepetersen, R. P. (2007), 'Modern Statistics for Spatial Point Processes*', *Scandinavian Journal of Statistics* **34**(4), 643–684.

Office for National Statistics (2019), 'Census geography - Office for National Statistics'.
URL: *https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography*

Ordnance Survey (GB) (2018), 'Points of Interest [CSV geospatial data], Scale 1:1250, Items: 670887'.
URL: *https://digimap.edina.ac.uk*

police.uk (2018), 'About | data.police.uk'.
URL: *https://data.police.uk/about*

police.uk (2019), 'Data downloads | data.police.uk'.
URL: *https://data.police.uk/data/*

PredPol (2019), 'PredPol Mission | About Us | Aiming to reduce victimization keep communities safer'.
URL: *https://www.predpol.com/about/*

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian processes for machine learning*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass. OCLC: ocm61285753.

Rengert, G. F. & Wasilchick, J. (2010), The Use of Space in Burglary, *in* 'Classics in environmental criminology', CRC Press, pp. 257–272.

Rousseau, J. & Mengersen, K. (2011), 'Asymptotic behaviour of the posterior distribution in overfitted mixture models: Overfitted Mixture Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 689–710.

Rue, H. & Held, L. (2005), *Gaussian Markov random fields: theory and applications*, number 104 *in* 'Monographs on statistics and applied probability', Chapman & Hall/CRC, Boca Raton.

Saatçi, Y. (2012), Scalable inference for structured Gaussian process models, PhD Thesis, Citeseer.

Sampson, R. J. & Groves, W. B. (1989), 'Community Structure and Crime: Testing Social-Disorganization Theory', *American Journal of Sociology* **94**(4), 774–802.

Sampson, R. J., Raudenbush, S. W. & Earls, F. (1997), 'Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy', *Science* **277**(5328), 918–924.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C. & Rue, H. (2014), 'Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008', *Environmental and Ecological Statistics* **21**(3), 531–563.

Shaw, C. R. & McKay, H. D. (1942), *Juvenile delinquency and urban areas : a study of rates of delinquents in relation to differential characteristics of local communities in American cities*, Chicago, Ill. : The University of Chicago Press.

Smith, K., Taylor, P. & Elkin, M. (2013), Crimes detected in England and Wales 2012/13, Statistical Bulletin 02/13, Home Office, London.

Stein, M. L. (1999), *Interpolation of spatial data: some theory for kriging.*, Springer, Place of publication not identified. OCLC: 968504419.

Taddy, M. A. (2010), '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.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic Geography* **46**, 234–240.

Tompson, L., Johnson, S., Ashby, M., Perkins, C. & Edwards, P. (2015), 'UK open source crime data: accuracy and possibilities for research', *Cartography and Geographic Information Science* **42**(2), 97–111.

Townsley, M., Birks, D., Bernasco, W., Ruiter, S., Johnson, S. D., White, G. & Baum, S. (2015), 'Burglar Target Selection: A Cross-national Comparison', *Journal of Research in Crime and Delinquency* **52**(1), 3–31.

Townsley, M., Birks, D., Ruiter, S., Bernasco, W. & White, G. (2016), 'Target Selection Models with Preference Variation Between Offenders', *Journal of Quantitative Criminology* **32**(2), 283–304.

A. Poisson regression model: excess of zeros, overdispersion

In this section we demonstrate that the standard Poisson regression (McCullagh & Nelder 1998) is not a suitable model for the London burglary point pattern.

Firstly, the dataset consists of areas with no buildings in it, e.g. parks, airports, which results in counts equal to zero due to structure rather than due to chance. This is further supported by the plot of the observed count and the corresponding histogram, both shown in figure 10. This phenomenon is often referred to as *excess of zeros.*

Secondly, we fit Poisson GLM with all four specifications of covariates to the 2015 burglary dataset, as described in section 3. Then we use the overdispersion test proposed in Cameron & Trivedi (1990), and implemented in the AER package (Kleiber & Zeileis 2008). For the standard Poisson GLM model, $\mathrm{Var}(\mathrm{y}_n) = \mu_n$. The overdispersion test uses it as the null hypothesis, where the alternative is $\mathrm{Var}(\mathrm{y}_n) = \mu_n + c \times g(\mu_n)$, where $g(\cdot)$ must be specified. For our test, we choose $g(\cdot) = 1$. Table 4 shows the estimated c values and the p-values for each estimate, given that null hypothesis is $c = 0$. The data clearly show the presence of overdispersion in all four models.

A.1. Poisson regression vs SAM-GLM

Figure 11 shows the scatter plot of expected vs observed counts for the Poisson regression model (SAM-GLM with $K = 1$) and the proposed model with $K = 3$. It is evident from the plot that adding extra components to the standard Poisson regression reduces the overdispersion issue.

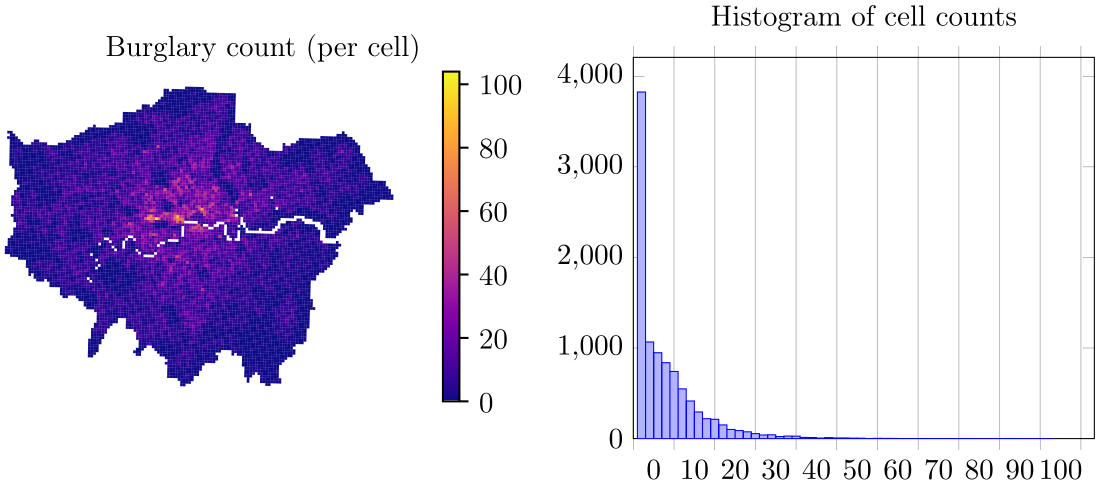


Figure 10. Observed count on the map (left) and the corresponding histogram (right) for the point pattern of burglary aggregated over the grid for the time period 1/2015-12/2015.

Table 4. Overdispersion test for Poisson GLM model.

Specification	c	p-value
1	1.905	2.2e-16
2	1.897	2.2e-16
3	1.910	2.2e-16
4	1.911	2.2e-16

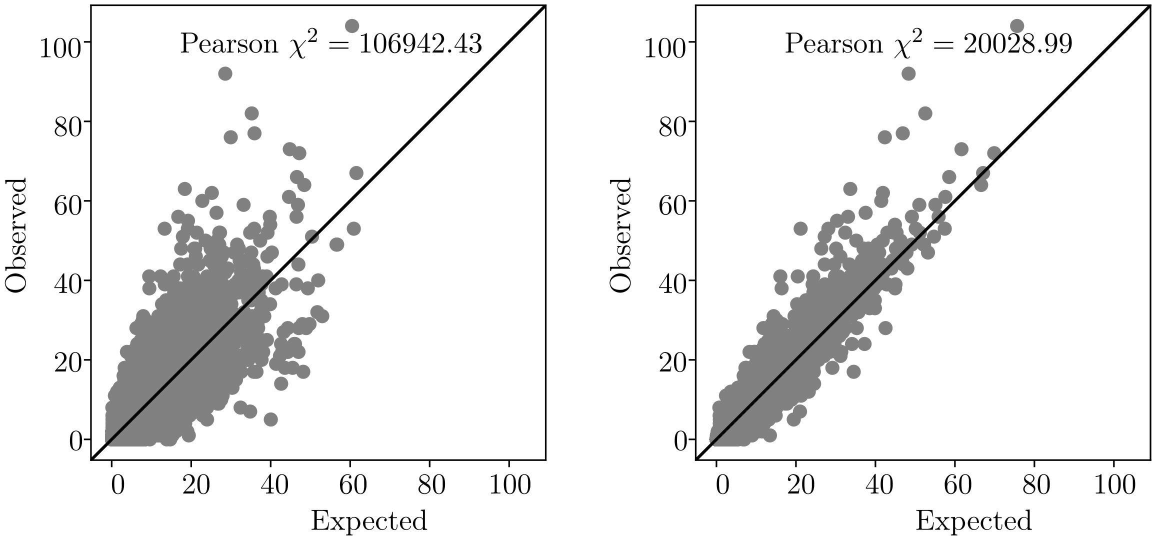


Figure 11. Scatter plot of predicted counts vs observed counts (training data) for the Poisson GLM model (left), and SAM-GLM K=3 (right). Blocking: MSOA, training data: 2015, using specification 4.

B. Log-Gaussian Cox process

Dicretising the spatial domain to a regular grid, the full Bayesian formulation of the model is given as follows:

$$\begin{aligned}
\mathrm{y}_n|\boldsymbol{\beta}, \mathbf{f}, \boldsymbol{X} &\sim \text{Poisson}\left(\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta} + \mathrm{f}_n)\right) & (10)\\
f(\cdot)|\boldsymbol{\theta} &\sim \mathcal{GP}\left(0, k_{\boldsymbol{\theta}}(\cdot,\cdot)\right) & (11)\\
\beta_j &\sim \mathcal{N}(0, \sigma_j^2) & (12)\\
\sigma_j^2 &\sim \text{InvGamma}(1, 0.01) & (13)\\
\boldsymbol{\theta} &\sim \text{weakly-informative log-normal prior}, & (14)
\end{aligned}$$

where $n = 1, \ldots, N$ is the index over the cells on the map, $j = 1, \ldots, J$ is the index over the covariates, $f()$ is a zero-mean Gaussian process with covariance function $k_{\boldsymbol{\theta}}(\cdot,\cdot)$, and hyperparameters $\boldsymbol{\theta}$, f_n is the value of $f(\cdot)$ in the centre of cell n, $\boldsymbol{X}_n$ is the vector of the covariates at cell n, and $\boldsymbol{\beta}_j$ is the jth regression coefficient with a scale hyperparameter σ_{kj}^2. A plain Poisson generalised linear model (GLM) formulation assumes no spatial correlation, i.e. $\mathrm{f}_n = 0$ for all n. Compared to the Poisson GLM model, LGCP allows for modelling the variation in the intensity that cannot be explained by the covariates $\boldsymbol{X}$.

In order to allow for Kronecker product factorisation of the covariance matrix of the Gaussian process, we specify $k_{\boldsymbol{\theta}}(\cdot,\cdot)$ as a product of two Matérn covariance functions, one for the easting (E) coordinate, the other for the northing (N) coordinate. Matérn covariance function is a standard choice in spatial statistics as it allows specifying smoothness of the function (Stein 1999). It is given as follows

$$k_{\text{Matern}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell} \right), \tag{15}$$

where ℓ is the characteristic lengthscale, ν is the smoothness parameter, and K_ν is a modified Bessel function (Rasmussen & Williams 2006). It can be shown that that the Gaussian processes with Matérn covariance functions are k-times mean-square differentiable if and only if $\nu > k$. Abramowitz & Stegun (1965) show that if ν is a half-integer, i.e. for an integer p, $\nu = p + \frac{1}{2}$, the covariance function becomes especially simple, giving

$$k_{\ell,\nu=p+1/2}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell} \right)^{p-i}. \tag{16}$$

For this reason, we set $\nu = 3/2$. The final covariance function, including the σ^2 parameter to control the range of $f()$ therefore becomes

$$k_{\boldsymbol{\theta}}((x_{\mathrm{E}}, x_{\mathrm{N}}), (y_{\mathrm{E}}, y_{\mathrm{N}})) \;=\; \sigma^2 k_{\ell,\nu=3/2}(x_{\mathrm{E}}, y_{\mathrm{E}}) \times k_{\ell,\nu=3/2}(x_{\mathrm{N}}, y_{\mathrm{N}}), \tag{17}$$

where $\boldsymbol{\theta} = [\sigma^2, \ell]^\top$.

B.1. Inference

To infer posterior distribution of the regression coefficients, $\boldsymbol{\beta}$, latent field $\mathbf{f}$, and its hyperparameters $\boldsymbol{\theta}$, we use a Hamiltonian Monte Carlo sampler. The scale parameters $\sigma_1^2, \ldots, \sigma_J^2$ are analytically integrated out (see equation 23 in the appendix). Due to positivity constraint of the hyperparameters, we sample from $\boldsymbol{\phi} = \log \boldsymbol{\theta}$ (applied component-wise). The density function of the joint posterior distribution we are sampling from is proportional to the product of likelihood and the priors, i.e.

$$p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) p(\mathbf{f}|\exp(\boldsymbol{\phi})) p(\boldsymbol{\beta}) p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) \prod_i \left| \frac{d}{d\phi_i} \exp(\phi_i) \right|. \tag{18}$$

To effectively use HMC sampler, log-likelihood of the posterior and its gradient need to be tractable. Thanks to the grid structure of our study region, we utilise Kronecker product structure that is present in the covariate matrix in $p(\mathbf{f}|\boldsymbol{\theta})$ if the covariance function $k_{\boldsymbol{\theta}}(\cdot,\cdot)$ is assumed to be a product of covariance functions, one per each dimension (For more details, see Saatçi (2012)). After expansion, the unnormalised log-density becomes

$$\begin{aligned}\log p(\mathbf{f},\boldsymbol{\beta},\boldsymbol{\phi}|\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{f},\boldsymbol{\beta}) + \log p(\boldsymbol{\beta}) + \log p(\mathbf{f}|\exp(\boldsymbol{\phi})) + \log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + \sum_i \phi_i + \text{const}_1 \\ &= \left(\mathbf{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \mathbf{y}^\top\mathbf{f} - \exp(\boldsymbol{X}\boldsymbol{\beta}+\mathbf{f})\right) + \log p(\boldsymbol{\beta}) \\ &\quad + \left(-\frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{\theta}}| - \frac{1}{2}\mathbf{f}^\top\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\mathbf{f}\right) + \log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + \sum_i \phi_i + \text{const}_1, \qquad (19)\end{aligned}$$

The gradients of the log posterior density w.r.t. quantities of interest are

$$\nabla_{\mathbf{f}} \log p(\mathbf{f},\boldsymbol{\beta},\boldsymbol{\phi}|\mathbf{y}) = (\mathbf{y} - \exp(\boldsymbol{X}\boldsymbol{\beta}+\mathbf{f})) + \left(-\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\mathbf{f}\right) \qquad (20)$$

$$\nabla_{\boldsymbol{\beta}} \log p(\mathbf{f},\boldsymbol{\beta},\boldsymbol{\phi}|\mathbf{y}) = \left(\boldsymbol{X}^\top\mathbf{y} - \boldsymbol{X}^\top\exp(\boldsymbol{X}\boldsymbol{\beta}+\mathbf{f})\right) + \nabla_{\boldsymbol{\beta}}\log p(\boldsymbol{\beta}) \qquad (21)$$

$$\begin{aligned}\nabla_{\phi_i} \log p(\mathbf{f},\boldsymbol{\beta},\boldsymbol{\phi}|\mathbf{y}) &= \frac{1}{2}\mathbf{f}^\top\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\mathbf{f} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\right) \\ &\quad + \nabla_{\phi_i}\log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + 1. \qquad (22)\end{aligned}$$

The expansion of un-normalised log-density of $\boldsymbol{\beta}$ and the gradients are derived in equation 24 and equation 25 below.

All operations involving $\boldsymbol{K}_{\boldsymbol{\theta}}$ can be sped up using Kronecker product factorisation. Given n^2 is the number of elements in the full matrix $\boldsymbol{K}_{\boldsymbol{\theta}}$, operations in equation 20 and equation 22 can be computed in $\mathcal{O}\left(n^{\frac{3}{2}}\right)$ time by utilising the Kronecker structure in matrix inversion and matrix-vector multiplication. For full details, see Saatçi (2012).

C. Model derivations

C.1. Beta prior

Given a vector of J independent random variables $\boldsymbol{\beta}$, of which each component is distributed as follows

$$\beta_j \sim \mathcal{N}(0,\sigma_j^2),$$
$$\sigma_j^2 \sim \text{InvGamma}(a,b).$$

Let $\Psi = \left(\sigma_1^2,\ldots,\sigma_J^2\right)^\top$, then the prior for the coefficients is given by integrating out the nuisance parameter Ψ

$$\begin{aligned}p(\boldsymbol{\beta}) &= \prod_j p(\beta_j) \\ &= \prod_j \int p(\beta_j|\Psi_j)p(\Psi_j)d\Psi_j \\ &= \prod_j \int \frac{1}{\sqrt{2\pi}}\Psi_j^{-1/2}\exp\left(-\frac{1}{2\Psi_j}\beta_j^2\right)\frac{b^a}{\Gamma(a)}\Psi_j^{-a-1}\exp\left(-\frac{b}{\Psi_j}\right)d\Psi_j \\ &= \prod_j \frac{b^a}{\sqrt{2\pi}\Gamma(a)}\int \Psi_j^{-a-\frac{1}{2}-1}\exp\left(-\frac{\frac{1}{2}\beta_j^2+b}{\Psi_j}\right)d\Psi_j \qquad (23) \\ &= \prod_j \frac{b^a}{\sqrt{2\pi}\Gamma(a)}\frac{\Gamma\left(\frac{1}{2}+a\right)}{\left(\frac{1}{2}\beta_j^2+b\right)^{\frac{1}{2}+a}}\end{aligned}$$

For the purposes of HMC, we derive both log-density and the gradient of log-density w.r.t. the each individual components. Log-density is given as

$$\log p(\boldsymbol{\beta}) = \sum_i - \left(\frac{1}{2} + a\right) \log\left(\frac{1}{2}\beta_i^2 + b\right), \tag{24}$$

from which the gradient is equal to

$$\frac{\partial \log p(\boldsymbol{\beta})}{\partial \beta_i} = \frac{(-\frac{1}{2} - a)\beta_i}{\frac{1}{2}\beta_i^2 + b}. \tag{25}$$

C.2. Conditional densities for SAM-GLM inference

The derivations below use the properties of the density function of the Dirichlet distribution and the following property of the Gamma function, $\Gamma(a+1) = a\Gamma(a)$.

C.2.1. Regression coefficients update

$$\begin{aligned}
p(\boldsymbol{\beta}|\alpha, \boldsymbol{X}, \mathbf{y}, \mathbf{z}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z})p(\boldsymbol{\beta}) \\
&\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N} p(\mathrm{y}_n|\boldsymbol{\beta}, \boldsymbol{X}, \mathrm{z}_n)\right\} \\
&\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathrm{y}_n|\boldsymbol{\beta}_k, \boldsymbol{X})^{I(\mathrm{z}_n=k)}\right\} \\
&\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N}\prod_{k=1}^{K} \left(\frac{\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)^{\mathrm{y}_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)}}{\mathrm{y}_n!}\right)^{I(\mathrm{z}_n=k)}\right\},
\end{aligned} \tag{26}$$

where $p(\boldsymbol{\beta})$ is expanded according to equation 23. For the purposes of Hamiltonian Monte Carlo, the gradient of the posterior distribution is analytically available.

C.2.2. GPs updates

The unnormalised joint posterior density of the K GPs and their hyperparameters is given as

$$\begin{aligned}
p(\boldsymbol{F}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) &\propto p(\mathbf{z}|\boldsymbol{F})p(\boldsymbol{F}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\propto \prod_{n=1}^{N} p(z_n|\boldsymbol{F}) \prod_{k=1}^{K} p(f_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k) \\
&\propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K}\exp(\mathrm{f}_{l,b[n]})}\right)^{I(z_n=k)} \prod_{k=1}^{K} p(f_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k),
\end{aligned}$$

where $p(f_k|\boldsymbol{\theta}_k)$ is the density function of the zero-mean multivariate Gaussian distribution with covariance matrix parameterised by $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}_k)$ is a suitable prior for the hyperparamers. The gradient of the joint posterior with respect to $\boldsymbol{F}$ and $\boldsymbol{\theta}$ are analytically available.

C.2.3. Mixture allocation update for spatially-dependent blocks

$$\begin{aligned}
p(\mathrm{z}_n = k|\mathbf{z}^{\bar{n}}, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}, \boldsymbol{F}) &= p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k)p(z_n|\boldsymbol{F}) \\
&\propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k)\frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K}\exp(\mathrm{f}_{l,b[n]})} \\
&= \prod_{k=1}^{K}\left(\frac{\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)^{\mathrm{y}_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)}}{\mathrm{y}_n!}\right)^{I(\mathrm{z}_n=k)}\frac{\exp(\mathrm{f}_{k,b[n]})}{\sum_{l=1}^{K}\exp(\mathrm{f}_{l,b[n]})}
\end{aligned}$$

C.2.4. Mixture allocation update for independent blocks

$$\begin{aligned}
p(\mathrm{z}_n = k|\mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}) &\propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\int p(\mathrm{z}_n|\boldsymbol{\pi}_{b[n]})p(\boldsymbol{\pi}_{b[n]}|\alpha, \mathbf{z}^{\bar{n}})d\boldsymbol{\pi}_{b[n]} \\
&\propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\int\prod_{k}\pi_{b[n],k}^{I(\mathrm{z}_n=k)}\frac{\Gamma(\sum_{j=1}^{K}B_{b[n],j})}{\prod_{j=1}^{K}\Gamma(B_{b[n],j})}\prod_{j=1}^{K}\pi_{b[n],j}^{B_{b[n],j}-1}d\boldsymbol{\pi}_{b[n]} \\
&\propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\frac{\Gamma(\sum_{j=1}^{K}B_{b[n],j})}{\prod_{j=1}^{K}\Gamma(B_{b[n],j})}\frac{\prod_{j=1}^{K}\Gamma(B_{b[n],j}+I(j=k))}{\Gamma(\sum_{j=1}^{K}B_{b[n],j}+I(j=k))} \\
&\propto p(\mathrm{y}_n|\mathrm{z}_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\frac{B_{b[n],k}}{\sum_{j=1}^{K}B_{b[n],j}} \\
&\propto \prod_{k=1}^{K}\left(\frac{\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)^{\mathrm{y}_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)}}{\mathrm{y}_n!}\right)^{I(\mathrm{z}_n=k)}\frac{c_{b[n]k}^{\bar{n}}+\alpha}{K\alpha+\sum_{j=1}^{K}c_{b[n]j}^{\bar{n}}},
\end{aligned} \tag{27}$$

where $B_{b,k} = c_{b,k}^{\bar{n}} + \alpha$, and $c_{b,k}^{\bar{n}}$ is the number of cells in block b other than cell n that are assigned to component k.

D. Dependence of blocks – extra plots

This section includes two plots related to the discussion of dependence of blocks in section 4.2.2. We compare the independent blocks version of our model with the variant that addresses the dependence via Gaussian random fields as described in section 2. The plots below show that considering dependence between the blocks can improve model predictions in some cases but it requires sampling from a high-dimensional distribution ($K \times B$), resulting in slow mixing.

Figure 12 compares smoothed histograms for samples of in-sample log-likelihood $p(\mathbf{y}|\boldsymbol{\phi})$ for both variants of the model when $K = 3$, with their out-of-sample counterpart using samples from $p(\tilde{\mathbf{y}}|\boldsymbol{\phi})$. While independent-blocks model performs better in-sample, the dependent-blocks model generalises better to out-of-sample data. However, for $K = 2$ and $K = 4$, the model with independent blocks has lower RMSE on out-of-sample data (reported in table 2).

Figure 13 shows the autocorrelation plot for the in-sample log-likelihood obtained from 50 000 samples that were thinned to 5000 for both variants to assess mixing performance. It is clear that successive samples obtained from the complex dependent-blocks model are more correlated to each other than for the case of independent blocks indicating slower mixing. Further, the inferences made using a Markov chain with high autocorrelation may lead to biased results.

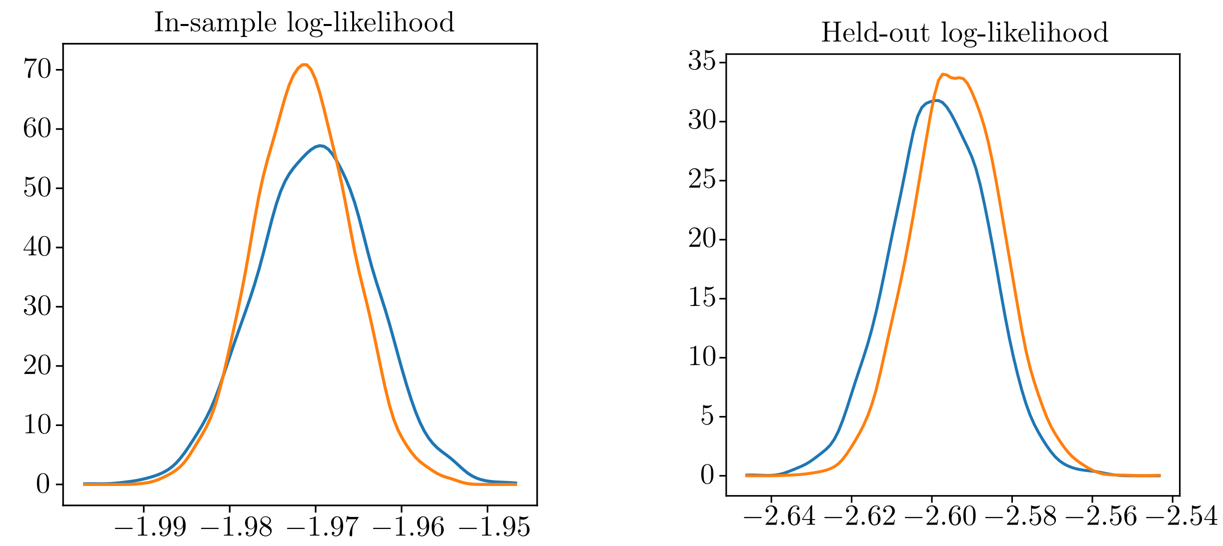


Figure 12. Smoothed histograms of log likelihood computed on in-sample counts (left), and out-of-sample counts (right) using the proposed model with dependent blocks (—), and independent blocks (—) when $K = 3$. Blocking: MSOA, training data: 2015, test data: 2016, model specification 4.

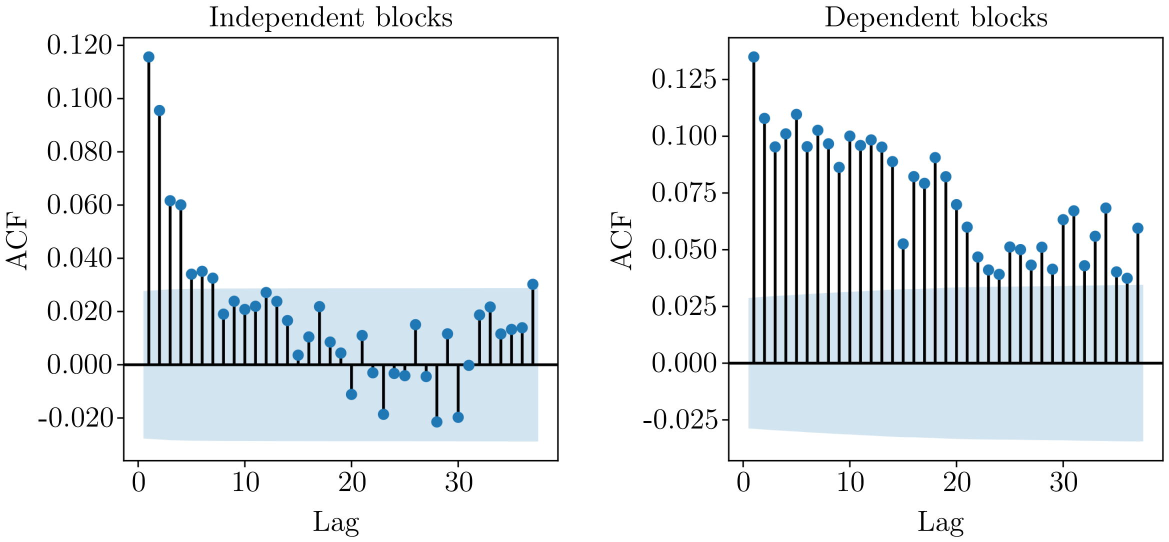


Figure 13. Autocorrelation plots for the samples of in-sample log-likelihood when $K = 3$. Blocking: MSOA, training data: 2015, model specification 4.