Exploring the usefulness of the INLA model in predicting levels of crime in the City of Johannesburg, South Africa

Abstract

Crime prediction serves as a valuable tool for deriving insightful information that can inform policy decisions at both operational and strategic tiers. This information can be used to identify high-crime areas, and optimise resource allocation and personnel management for crime prevention. Traditionally, techniques such as the Poisson model and regression analysis have been widely used for crime prediction. However, recent statistical advancements have introduced Integrated Nested Laplace Approximations (INLA) as a promising alternative for spatial and temporal data analysis. This study focuses on crime prediction using the INLA model. Specifically, the first-order autoregressive model under the INLA modelling framework is employed on longitudinal data for crime predictions in different regions of the City of Johannesburg, South Africa. The model parameters and hyperparameters considering space and time are estimated through the INLA model. In this work, the suitability and performance of the INLA model for crime prediction is assessed, which effectively captures spatial and temporal patterns. This study contributes to research by first introducing a novel approach for South African crime prediction. Secondly, it develops a model using no demographic information other than clustering attributes as an exogenous variable. Thirdly, it quantifies prediction uncertainty. Finally, it addresses data scarcity through demonstrating how INLA can provide reliable crime predictions, where conventional methods are limited. Based on our findings, the INLA model ranked areas by crime levels, obtaining a 29.3% Mean Absolute Percentage Error (MAPE) and 0.8 R^2 value for crime predictions. These findings and contributions presents the potential of INLA in advancing evidence-based decision-making for crime prevention.

Keywords: Integrated Nested Laplace Approximation, Bayesian Inference, Crime Statistics, Crime Prediction

1 Introduction

Ranked among the countries with the highest crime rates globally (Institute for Economics and Peace 2023), South Africa grapples with pervasive levels of criminal activity, further compounded by a notably low ratio of police officers to population.

According to estimates by the United Nations (UN), the average police officer-to-population ratio globally is 342 police officers per 100,000 people. In contrast, South Africa reports a lower ratio, with only 240 police officers per 100,000 people (SAPS 2023). As a result, police visibility suffers, making it extremely difficult to combat crime, especially offences that could be controlled through visible policing. South Africa's alarming crime rate and low police officer-to-population ratio underscores the urgency and motivation for developing effective methodologies to predict and mitigate crime. Therefore, this work implements predictive modelling using crime data publicly available and published quarterly by the South African Police Service (SAPS). By focusing on South African crime data, this study aims to develop a predictive model to better suit the unique characteristics and challenges of this specific context. This approach is significant because it acknowledges the distinct nature of South African crime data compared to other international datasets typically used in predictive crime modelling. Furthermore, leveraging the advancements in science and technology can prove to be pivotal in such cases. This research thus intends to develop a tool that uses both statistics and geographic information systems (GIS) to provide crime analytics and one-quarter-ahead crime forecasts at the police station area level. The predictive capability of the tool would enable the police to be proactive rather than reactive in their approach to combating crime. By adapting and refining predictive models tailored to South Africa, this work seeks to enhance the accuracy and reliability of crime prediction methodologies and optimise police resource allocation in the country. This optimisation of police resource allocation not only aids law enforcement and policymakers in combating the country's severe crime situation but also revolutionises policing strategies by promoting well-informed, evidence-based practices.

Predictive policing is one of the applications of analytical techniques in which researchers and law enforcement agencies work together to combat crime by developing tools that predict crime, resulting in the more precise and effective use of law enforcement resources. For instance, Borges et al. (2018) developed a crime prediction framework that contributes to predictive policing initiatives by incorporating patrols in the most likely predicted criminal areas. Blanes I Vidal and Mastrobuoni (2018) also carried out a large, scientific-based study on visible policing in the UK. Their findings revealed that although simply increasing police presence through street patrols was able to deter crime to some extent, traditional patrolling methods did not yield significant long-term crime prevention outcomes. Consequently, integrating science and technology into policing operations ensures resource allocation optimisation, maximising their impact on crime prevention and law enforcement efforts.

In this work, we explore the use of the INLA model as our model of interest to predict levels of crime in the City of Johannesburg (CoJ). We aim to identify potential high-crime areas within the CoJ to facilitate targeted police resource allocation. The crime dataset that we are using has sparse data in terms of both temporal and spatial resolution, aggregated at the police station boundary level. This presents a challenge for crime prediction within a context of scarce data availability. The study thus also

explores the potential of advanced statistical techniques like INLA to produce reliable crime predictions despite data scarcity. The subsequent sections of the paper will be structured as follows: Firstly, the literature review will discuss applications of INLA and other approaches to crime prediction such as spatial regression and Bayesian networks. Next, we describe the data used, providing an exploratory data analysis. We then discuss the methodology of using INLA, describing the model, its mathematical underpinnings, as well as the evaluation methods used for assessing the INLA model. Afterwards, we present the results of using the INLA model, including its goodness-of-fit, predictions for the various regions of Johannesburg, and its performance based on evaluation metrics such as MAPE. Finally, we discuss this work, its limitations, and conclude with the significance of the results of the INLA model in the context of crime prediction in South Africa.

2 Literature review on approaches to crime prediction

Crime prevention and public safety are crucial concerns for societies around the world. Law enforcement agencies and policymakers strive to develop effective strategies to combat crime and allocate resources efficiently. In recent years, there has been a growing interest in utilising mathematical and statistical methods to predict crime occurrences and assist in proactive law enforcement efforts (Kang & Kang 2017; Khan, Ali, & Alharbi 2022). These methods leverage the power of data analysis, modelling, and predictive analytics to identify patterns, understand the underlying factors, and forecast crime rates and hot-spots. In this literature review, we introduce Integrated Nested Laplace Approximation (INLA), then explore different methodologies used in crime prediction, including approaches such as Bayesian networks and regression. We look at how previous research have similarly used historical crime data for the same motivation of crime prediction. We also note the limitations of each method, how INLA may address these and the scarcity of similar studies conducted in the South African context.

The Integrated Nested Laplace Approximation (INLA) is a Bayesian inference alternative to the computationally intensive Markov Chain Monte Carlo (MCMC) methods. INLA enables fast approximate inference for latent Gaussian models. Boqué, Saez, and Serra (2022) highlighted that, one of the INLA advantages is that it avoids long computing time compared to the MCMC approach. Furthermore, Boqué et al. (2022) have successfully applied the logarithmic-Gaussian spatio-temporal model with INLA to predict weekly burglaries, which has yielded notable results. This model takes into account the latent risk of burglary in both space and time, and was able to identify the spatial correlation in the range of distances. Consequently, each burglary increases the probability of subsequent burglaries both close to and further away from the initial incidence. Their findings advocate for the use of INLA as a model proficient in studying or forecasting near-repeat victimization incidents. Vicente, Goicoa, and Ugarte (2023) proposed multivariate Bayesian spatio-temporal P-splines models fitted using INLA to study various types of violence against women in India. The model

was effective in showing spatial patterns of different types of violence against women and in identifying high-risk crime in specific areas (Vicente et al. 2023).

Regarding other approaches to crime prediction, we first consider Bayesian networks, which are probabilistic graphical models that can represent the relationships between variables and provide predictions based on available evidence. In the context of crime prediction, Bayesian networks can capture dependencies between different factors (e.g., demographics, past crime data, environmental factors) and estimate the likelihood of future crime occurrences (Liao, Wang, Li, & Qin 2010). A Bayesian-based crime prediction model was developed by Liao et al. (2010) using geographic data and victim attributes based in Baiyin city, China, over several months. The authors predicted the next crime site chosen by a serial offender, with a particular focus on geographical factors. They developed a geographic profile, which represents the probability distribution of crime events, using a discrete distance decay function. Finally, Bayesian learning theory was adapted with geographic characteristics to precisely forecast crime levels in regions of interest. However, a limitation of Bayesian networks is that they may struggle with rare or low-frequency events. Crime prediction often involves identifying and predicting rare or emerging criminal activities. Bayesian networks may face challenges in accurately capturing and predicting these rare events due to the limited availability of training data for such events. Another limitation relevant to this study in the context of Bayesian networks is that they typically predict events or outcomes rather than counts over a specific area (Hu, Zhu, Duan, & Guo 2018). INLA can handle rare events more effectively by leveraging hierarchical modelling techniques, which enable the borrowing of strength across different spatial and temporal units. Moreover, INLA's ability to model counts over specific areas makes it well-suited for crime prediction tasks that involve predicting crime counts at police station area or grid cell levels.

Regression analysis is another statistical technique used for crime prediction, that explores the relationship between a dependent variable (such as crime rates) and one or more independent variables (such as demographic variables, socioeconomic factors, or past crime data). Specifically, spatial regression analysis is a statistical technique that combines the components of traditional regression analysis with spatial analysis to simulate correlations between variables, while taking into consideration spatial dependencies in the data. The regression model known as Ordinary Least Squared (OLS) and the spatial technique known as Spatial Autoregressive (SAR) were employed by Ahmar, Adiatma, and Aidid (2018) to predict and model crime. They also incorporated the Lagrange Multiplier (LM) to detect the existence of spatial dependency. Additionally, Chainey, Tompson, and Uhlig (2008) conducted a study that focused on hot-spot mapping as a fundamental method of predicting crimes. However, a limitation of traditional spatial regression analysis is the computational burden and instability when dealing with large datasets or complex spatial structures (Urdangarin, Goicoa, & Ugarte 2023). INLA overcomes this computational burden by employing a computationally efficient and accurate Laplace approximation. This approximation method allows for faster computation while still providing accurate

posterior estimates, making it particularly suitable for large and complex spatial models (Urdangarin et al. 2023). Moreover, INLA does not require a large dataset as required for regression, making it suitable for this context where a large dataset is unavailable in South Africa.

The Poisson model is one of the regression models for count data that has been widely used for crime prediction and analysis due to its applicability in capturing the characteristics of crime incidents. Muchika, Ngunyi, and Mageto (2020) applied Generalised and Quasi Poisson regression to burglary crime data in Kenya. In their findings, the Generalised Poisson provided nearly an excellent fit for predicting burglary incidents in Nairobi, Kenya as compared to Quasi Poisson. Their study further revealed that for under-dispersed count data, the Generalised Poisson performs better than both the standard and Quasi Poisson.

However, conventional models like Poisson regression and OLS have limitations, such as the strict assumption of data following a Poisson distribution and the inability to handle spatial and temporal dependencies effectively. For instance, the Poisson model cannot accurately forecast crime counts in regions with no reported crime, limiting its practical applicability (Gordon 2010; Poyton, Varziri, McAuley, McLellan, & Ramsay 2006). INLA offers a more flexible modelling framework that can accommodate various distributional assumptions and handle spatial and temporal structures in the data. Its hierarchical modelling approach allows for capturing complex relationships between crime variables and potential factors while addressing non-stationarity and spatial dependencies. Additionally, INLA's spline-based modelling can improve forecasting accuracy by extrapolating basis functions beyond the region of estimation, enabling more reliable predictions even in regions with sparse or missing data. Moreover, approaches such as Poisson often fail to quantify uncertainties associated with model parameters, such as regression coefficients (Gourieroux, Monfort, & Trognon 1984). INLA addresses this limitation by adopting a probabilistic framework, generating posterior distributions that reflect uncertainties in crime prediction models (Muff, Riebler, Held, Rue, & Saner 2015). Through these distributions, credible intervals are calculated, providing a measure of uncertainty for estimated parameters and facilitating informed decision-making. Although the standard Poisson model performs better with included basis functions, their subjective selection may lead to misspecification (Marra & Radice 2010). INLA overcomes this by automating basis function selection through model comparison and Bayesian model averaging, improving the fitting process, and capturing complex predictor-crime count relationships (Louzada, Nascimento, & Egbon 2021). Additionally, INLA can estimate model parameters and uncertainties even in regions with limited or no reported crime data, addressing the Poisson model's limitation in extrapolation. Furthermore, INLA provides a framework to model over-dispersed count data and incorporate additional spatial or temporal factors, enhancing the representation of crime patterns (Louzada et al. 2021).

Overall, while INLA has been employed for crime prediction internationally, it is

yet to be tested in the South African context as seen by the majority of literature originating internationally (Boqué et al. 2022; Vicente et al. 2023).

3 Data

3.1 Data Description

Fig. 1: Police station boundaries in the City of Johannesburg.

The reported crime data used in this study was collected in South Africa over quarters (time) from 2017 to the fourth quarter of 2021, and are indexed by both time and space (spatio-temporal). The data were split into training and test sets. Due to limited data, we allocated 95% to training and 5% to testing to maximise the training data and build a robust model. Crime data from 2017 to the third quarter of 2021 were used as the training data while the fourth quarter of crime data were used as test data. The training data were used to train the model and the test data were used to test the model. Recall that we were operating in a data-scarce environment, with publicly available data sets containing only aggregated information (spatially and temporally). For this reason, the application of techniques such as time series analysis and machine learning were not feasible. These crime statistics include crime counts for different crime categories, and are recorded in every police station in South Africa, but in this study, only the data from the City of Johannesburg (CoJ) were used for our model. In addition, only certain of the SAPS crime categories (those more related to visible policing) were selected for later analysis and these categories are listed in Table A1 in the Appendix. Figure 1 above illustrates police station boundaries, or areas, for all police stations located within CoJ. Additionally, by aggregating crime

counts across the crime types, we circumvent the challenge of temporal sparsity while preserving the predictive accuracy essential for area-level analysis aimed at targeting crime hot-spots.

As mentioned, the crime data is publicly available and published quarterly by the South African Police Service (SAPS). The polygons in Figure 1 represent 43 demarcated SAPS areas around police stations that encompass a variety of socioeconomic and crime-type characteristics. With regards to data subjectivity, the crime data used faces inherent biases due to its reliance on police reports. Human reporting introduces subjectivity, potentially leading to under-reporting or over-reporting based on various factors. Additionally, the accuracy of reported crime locations may be compromised, as incidents are recorded at police stations or hospitals rather than the actual crime scene. Moreover, certain crimes are often under-reported, particularly for offences such as theft and vandalism, which are often only reported for insurance purposes. We thus acknowledge this inherent bias in the data and note its limitations for this study. In our study, we mitigate bias inherent in the crime data by capturing uncertainties. Specifically, we calculate the 2.5th, 50th (posterior mean), and 97.5th percentiles of the posterior distribution obtained from our predictive model. These percentiles provide valuable insights into the uncertainty associated with our predictions, allowing us to account for biases in the data, such as under-reporting and inaccuracies in crime location, as well as model uncertainties. By considering a range of possible scenarios rather than relying solely on point estimates, we enhance the robustness of our analysis.

3.2 Exploratory analysis of data

As mentioned above, crime counts were only available at a police station area level. In order to visualise these data spatially, the map given in Figure 2 was generated, illustrating the average quarterly crime counts per police station across the selected crime categories listed in Table A1. Spatial variations in crime counts can be observed across these different police station areas.

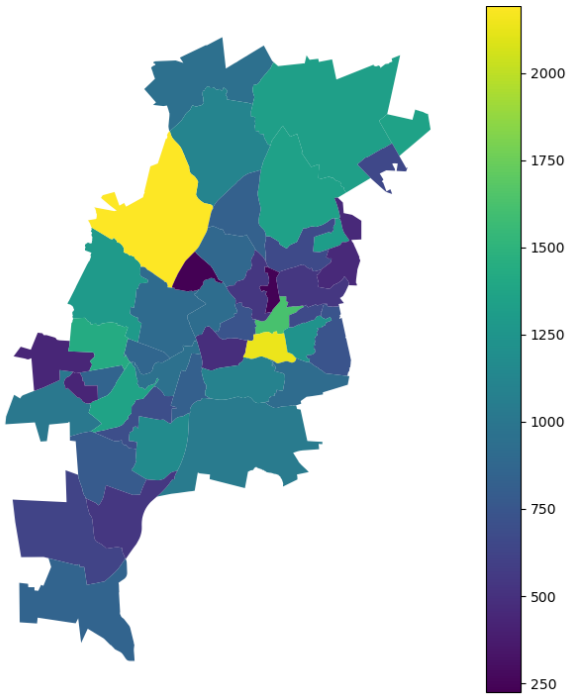


Fig. 2: Plot of average quarterly crime counts per police station area in CoJ.

Time-based trends of crime were observed by plotting crime counts per crime type over time. An example of these trend graphs is given in Figure 3. This plot includes the trends of specific burglary and theft crime categories and indicates a general downward trend in these crime types.

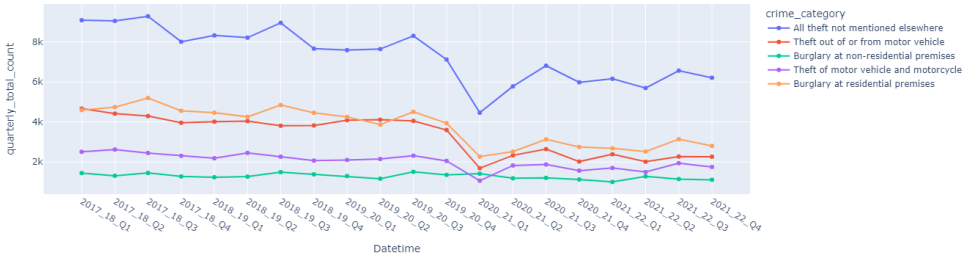
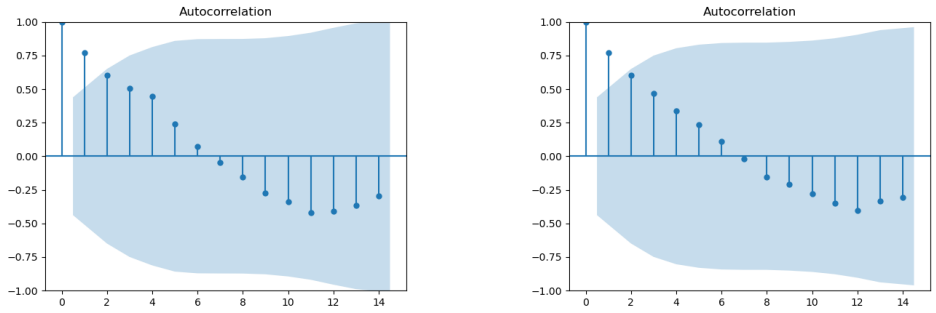


Fig. 3: Quarterly trend of CoJ crime counts for thefts and burglaries.

Before starting the predictive modelling, further exploratory analysis was performed on the SAPS data to determine whether there was any temporal or spatial autocorrelations in the data. To test for temporal autocorrelation in the data, the Ljung-Box test was used and if significant then the Autocorrelation Function (ACF) plot was drawn (see Armstrong (2001) for further details on these measures). Many of the crime types exhibited serial correlation using both the Ljung-Box test and the ACF plot. For example, Figure 4 indicates the strong time-based autocorrelation for the crimes "Burglary at residential premises" and "Theft out of motor vehicle" from lag1. These lag1 autocorrelations correspond to the trends seen in Figure 3.

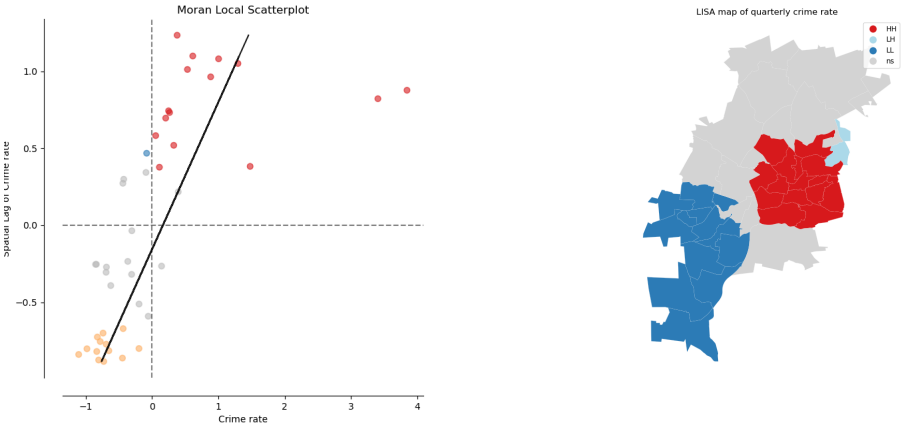


(a) (b)

Fig. 4: ACF plot of a) Burglary at residential premises and b) Theft out of motor vehicle.

Spatial autocorrelation can be defined as the presence of systematic spatial variation and in order to test for positive spatial autocorrelation, i.e. the tendency of areas close together to have similar values (Haining 2001), Moran's I (Dubé & Legros 2014), together with Local Indicators of Spatial Association (LISA) maps (Jesri et al. 2021), were selected as the tools for assessing spatial autocorrelation. The Moran's I statistic for these data of overall crime averages per police station was 0.5 (p-value: 0.001) and therefore indicated significant spatial autocorrelation overall. The Moran's local I scatter plot and the LISA map are given in Figure 5.
The LISA map clearly shows the spatial separation of different crime rates, with high crime rates being clustered around the central Johannesburg region and low crime rates being clustered together down in the Southwest area of the city region.

The significant spatial autocorrelation observed in Figure 5 indicates the requirement of including a spatial element in the crime prediction modelling and the evidence of temporal autocorrelation in the data for a number of the crime types, as illustrated in Figure 4 above, suggested the need for introducing a prediction model that included an autoregressive (AR) component. This combined evidence for both spatial and temporal autocorrelation resulted in the need to consider a spatio-temporal modelling



(a) (b)

Fig. 5: Plots of the a) Moran's local I statistic and b) LISA map.

technique that would be appropriate for the available data. This spatio-temporal modelling technique is discussed in the next section.

4 Methodology

4.1 Definition and Implementation of INLA

This study predicts regions in which crime is most likely to occur so that preventative measures may be employed in the most efficient and effective way. Intuitively, the immediate past values of a variable should have a better forecasting ability to predict near future values, therefore, the simplest autoregressive model would give the most recent observed outcome of the time series a higher weight in predicting future values. The time series of crime counts (y_t) can be modelled with a first-order autoregressive model (AR(1)), where 1 indicates the order of autoregression represented in Equation (1) as

$$y_t = \theta_0 + \theta_1 y_{t-1} + \varepsilon_t \tag{1}$$

where:

- y_t is the count of crime at each subplace at quarter t
- y_{t-1} is the count of crime in the previous quarter
- θ_1 is the coefficient for y_{t-1}, representing the relationship between crime count at the current quarter and crime count at previous quarter. The value of θ_1 will always be 1 or –1 for the series to satisfy the assumption of stationarity
- ε_t is the error term at time t. This represents the difference between the period t value and the predicted value using the model.

In this model, the previous crime counts in a specific quarter are used to predict the crime counts in the next quarter. The statistical model used to predict crime relies on the assumption of some degree of dependence in time and between locations. The multivariate Poisson-based models are a natural match to spatio-temporal time series of counts, and these have been employed in various applications. Bayesian-based models, which enable the quantification of uncertainties around predictions, have been applied in spatio-temporal data and performed well. Under the Bayesian modelling paradigm, when the posterior distribution is not available in a closed form, this necessitates resorting to other numerical methods such as the Markov Chain Monte Carlo (MCMC), for its estimation. The aim of the Bayesian model is to estimate the joint posterior distribution, which relies on Bayes theorem, shown in Equation (2) as

$$\pi\left(\theta|y\right) = \frac{\pi\left(y|\theta\right)\pi\left(\theta\right)}{\pi\left(y\right)} \tag{2}$$

where:

- $\pi\left(\theta|y\right)$ represents the posterior distribution of the model parameters θ given the observed crime counts y.
- $\pi\left(y|\theta\right)$ denotes the likelihood function, indicating the probability of observing the crime counts y given the model parameters θ.
- $\pi\left(\theta\right)$ signifies the prior distribution of the model parameters, representing our initial beliefs or knowledge about the parameters before observing the data.
- $\pi\left(y\right)$ serves as the marginal likelihood function of crime occurrence, acting as a normalisation constant to ensure that the posterior distribution integrates to one.

The posterior distribution $\pi\left(\theta|y\right)$ of crime levels is multivariate and is only available in closed form from a few models because the marginal likelihood $\pi\left(y\right)$ is difficult to estimate. Hence, in practice, the posterior distribution is estimated without computing the marginal likelihood. For this reason, Bayes' theorem is often expressed as shown in Equation (3):

$$\pi\left(\theta|y\right) \propto \pi\left(y|\theta\right)\pi\left(\theta\right) \tag{3}$$

This demonstrates the proportionality between the posterior distribution of the model parameters θ given the data y, and the product of the likelihood of the data given the model parameters θ and the prior distribution of the model parameters θ. This means that the posterior distribution can be estimated by rescaling the product of the likelihood and the prior so that it integrates to one. Bayesian inference relies on MCMC methods, which are computationally expensive, as the model parameters of the posterior distribution are often found in spaces of high dimension. Instead of estimating a highly multivariate joint posterior distribution $\pi\left(\theta|y\right)$, the Integrated Nested Laplace approximation (INLA) is used. The goal of INLA is to obtain the approximations of the univariate posterior distributions $\pi\left(\theta_i|y\right)$ represented in Equation (4) by,

$$\pi\left(\theta_i|y\right) = \int_1^{\dim(\theta)} \pi\left(y|\theta\right) d\theta_{-i} \ , \tag{4}$$

where:

- θ_i represents a specific parameter of interest within the set of parameters θ.
- θ_{-i} represent all the other parameters in the set θ, excluding θ_i. This includes all parameters except the specific parameter of interest.
- i ranges from 1 to $dim(\theta)$.

The integral in Equation (4) indicates that we're integrating over all possible values of the other parameters (θ_{-i}), while keeping θ_i fixed. This integration helps us to compute the posterior distribution of θ_i given the observed data y.

Regarding the lack of bounds for the integral, in the context of INLA, the integral is typically performed over the entire parameter space. This means that we are considering all possible values of the parameters within their feasible range. The lack of explicit bounds in the equation doesn't mean that the integral is unbounded; rather, it is understood to be performed over the appropriate range of values for each parameter.

The posterior marginal distribution of each element of θ can be obtained by integrating out the remainder of the parameters. The integrals of this type can be conveniently approximated using numerical integration methods and the Laplace approximation (Tierney & Kadane 1986). INLA makes Bayesian inference faster since, instead of aiming at estimating the joint posterior distribution of the model parameters, it focuses on individual marginal distributions of the model parameters. INLA estimates the fixed and random effects of the model and uses complex covariance structures to improve crime predictions. As the entire crime distribution is predicted, the uncertainty surrounding crime projections are quantified. In this study, the modelling approach limited the inclusion of exogenous variables, since police intervention in areas with high predicted levels of crime could influence the crime predictive ability of demographic variables.

In this study, the separability assumption was considered, which asserts that the space and time correlation can be modelled separately, allowing time autocorrelation to be captured in the autoregressive model while the spatial correlation is examined using the Besag-York-Mollié (BYM) model. Morris et al. (2019) defines the BYM model as a lognormal Poisson model which includes both an ICAR component for spatial auto-correlation and an ordinary random-effects component for non-spatial heterogeneity. This model was selected for our work because it assumes spatial correlations in the data, indicating that observations from neighboring areas are likely to be more similar to each other than areas that are further away (Moraga 2019). The first-order autoregressive model within the INLA framework was fitted using historical crime counts to predict future crime counts in areas within the City of Johannesburg. Before fitting the spatio-temporal model, we first computed the adjacency structure of the sub-places within the City of Johannesburg using shapefile datasets. Additionally, we opted for a separable model in this study. In particular, the spatial effect was modelled using an Intrinsic Conditional Auto-Regressive (ICAR) model and the temporal

trend using an Autoregressive (AR1) latent effect, and the vague priors were used to avoid over-fitting. According to Besag (1974), when areal data are assumed to have a spatial structure such that observations from neighbouring regions exhibit higher correlation than distant regions, this correlation can be accounted for by using the class of spatial models called "CAR" (Conditional Auto-Regressive). This calculation was done using a function `poly2nb()` in R (R Core Team 2021). In the simplified form, the model fitted is:

```
model <- inla(crime ~ Quantile +
                  f(quarter, model = "ar1") +
                  f(Areas, model = "besag", graph = adj.mat))
```

where:

- `besag` is the Besag spatial model.
- `adj.mat` is the adjacency structure of the subplaces in the City of Johannesburg.
- `ar1` is the autoregressive model of order 1.
- `crime` is the dependent variable.
- `Quantile` is the fixed effect.
- `F(quarter,model=ar1)` is the temporal term of the model (time series).
- `f(Areas, model = "besag", graph = adj.mat)` is the spatial component of the model.

The complete fitted model with the poly2nb() function used can be found in the code's GitHub repository[1].

4.2 Evaluating the INLA Model

The appropriateness and usefulness of the INLA model was evaluated by assessing its Goodness-of-fit. This included calculating the Conditional Predictive Ordinate (CPO), Probability Integral Transform (PIT) values, as well as the R-squared and Kolmogorov-Smirnov (KS) statistic. Subsequently, the performance of the model was evaluated through the use of the Mean Absolute Percentage Error (MAPE), and an analysis of the resultant actual vs predicted crime counts graph with a 95% confidence interval.

Goodness-of-fit serves as a reliable metric for evaluating the effectiveness and appropriateness of our model by evaluating how well it fits the observed data. This evaluation provides insights into the model's ability to capture the complex patterns and variations present in crime data. To assess the Goodness-of-fit for the INLA model, firstly, the Conditional Predictive Ordinate (CPO) and Probability Integral Transform (PIT) values were calculated as follows:

Conditional Predictive Ordinate (CPO):

$$CPO_i = \frac{1}{p(y_i|\mathbf{y}_{-i})} \tag{5}$$

[1] https://github.com/CSIR-CoJ-Crime-Project/COJ-INLA-Predictions.git

where:

- y_i denotes the i-th observed crime count.
- $\mathbf{y}_{-i}$ refers to the vector of all observed crime counts excluding the i-th observed crime count.
- $p(y_i|\mathbf{y}_{-i})$ is the conditional predictive distribution of the i-th observation given the remaining data points.

Probability Integral Transform (PIT):

$$PIT_i = \int_{-\infty}^{y_i} p(y_i|\mathbf{y}_{-i})dy_i, \tag{6}$$

where:

- y_i denotes the i-th observed crime count.
- $\mathbf{y}_{-i}$ refers to the vector of all observed crime counts excluding the i-th crime observation.
- $p(y_i|\mathbf{y}_{-i})$ is the conditional predictive distribution of the i-th observation given all crime observations excluding the i-th crime observation.

Next, the coefficient of determination, often denoted as R^2, was calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

where:

- y_i is the actual crime count value at quarter i,
- $\hat{y}_i$ is the predicted value at quarter i,
- $\bar{y}$ is the mean of the actual crime count values,
- n is the total number of quarters.

We also included the Kolmogorov-Smirnov (KS) statistic, defined in Equation 8 below, to supplement the reliability and appropriateness of our model's predictions, by further evaluating the overall Goodness-of-fit of our model's predictions. The KS statistic measures the difference between the cumulative distribution functions of predicted and actual crime counts. The Kolmogorov-Smirnov test statistic is given by:

$$D_n = \sup_x |F_n(x) - F(x)| \tag{8}$$

where $F_n(x)$ is the empirical distribution function of the sample and $F(x)$ is the theoretical cumulative distribution function.

Finally, the Mean Absolute Percentage Error (MAPE) is a commonly used metric to evaluate the ability of forecasting models. It reflects the accuracy of a model's

predictions in comparison to its actual outcomes. The MAPE is calculated as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100\% \tag{9}$$

where:

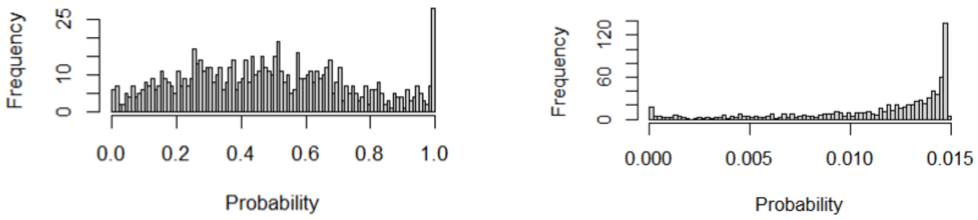
- A_i is the actual value,
- F_i is the forecasted (predicted) value,
- n is the total number of observations.

5 Results

In this section, we provide the results after applying the INLA model to predict future crime counts in the 43 police station areas in CoJ during the fourth quarter of the 2021 financial year, based on historic quarterly records.

5.1 Goodness-of-fit of the INLA model

If the model accurately predicts the crime levels, then the PIT values should be approximately uniformly distributed. In Figure 6(a) below, the graph for the PIT shows an approximate uniform distribution and thus predicts fairly well. Next, the model suggests that an observation with a small CPO value is unlikely. Based on Figure 6(b), there are no CPO values that are considerably smaller than others hence, with respect to the model, none of the observed values would be considered unlikely.



(a) PIT distribution. (b) CPO distribution.

Fig. 6: Goodness-of-fit for INLA model showing PIT and CPO distributions. (a) represents the PIT distribution and (b) represents the CPO distribution.

The R^2 value of the model was 0.8, indicating that the model explains 80% of the variability in the observed crime data. This shows that the INLA model fits the crime data well, indicating that the predictions closely match the observed data. Thus the model accurately represents the observed data and captures the underlying patterns and relationships. Finally, the KS statistic (D = 0.18605) is relatively small, suggesting that the predicted and actual distributions are similar. This affirms the

adequacy of the model in capturing the underlying distribution of the data.

Therefore, the use of the INLA model to predict crime count in regions of the City of Johannesburg in quarter four in 2021 was appropriate and justifiable.

5.2 Predictive performance of the INLA model

Figure 7 provides a comparison between the predicted average crime counts and the actual crime counts of the police stations during the fourth quarter of 2021 financial year. A scatter plot of these results is shown together with a 95% confidence interval.The 95% confidence interval indicates that there is a 95% probability that the true mean falls within the calculated interval. It was calculated by taking the mean and adding and subtracting the margin of error. This margin of error was determined by multiplying the standard error by the critical value corresponding to a 95% confidence level from the posterior distribution. Table A2 and Figure A1 in the Appendix provide the predictions at 2.5, 50 (posterior mean) and 97.5 percentiles of the posterior distribution. It is found that the model performed relatively well, with some variance between the actual and predicted values and a few stations either under-predicted or over-predicted. This is evident as the plot shows that the points are dispersed around the diagonal line, with closer proximity to the line indicating smaller residuals and thus a stronger fit between predicted and actual crime values. This observation underscores the model's effectiveness in accurately representing the data.

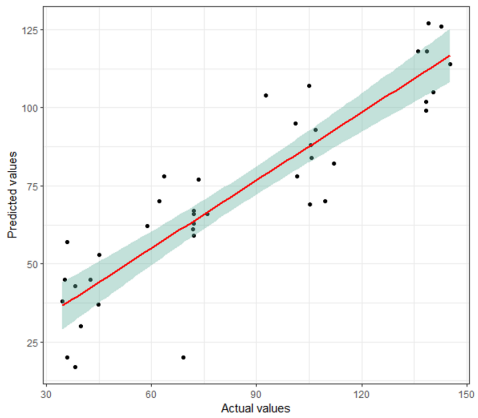


Fig. 7: Actual crime counts versus predicted crime counts for the fourth quarter of 2021 in CoJ, with a 95% confidence interval.

The crime forecasts are displayed spatially in Figure 8 below using a colour scale to show the prevalence of crime in each police station area. The areas shaded in yellow indicate high levels of crime while the dark purple show low levels of crime. Figure 8 indicates that areas such as Midrand, Sandton, Douglasdale, Roodepoort, Dobsonville and Moroka were predicted to have high levels of crime in quarter 4 of 2021. Areas adjacent to each other do not appear to be affected similarly by crime. We thus were able to identify the high-crime areas using the INLA model. The MAPE calculated from

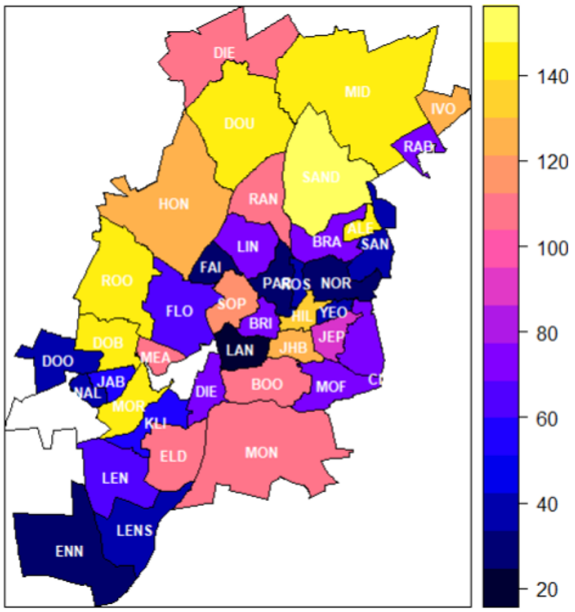


Fig. 8: Spatial representation of crime predictions in police station areas within CoJ.

the predictions was 29.3%, indicating that the predictions were, on average, 29.3% different from the actual values. We also observed in Figure 7, where actual crime counts are plotted against predicted crime counts, that the model performs reasonably well, with data points scattered around the diagonal line on the two-dimensional Cartesian plane.

6 Discussion

The evaluation of the INLA model's Goodness-of-fit and predictive performance holds significant implications for crime prediction in the CoJ. The Probability Integral Transform (PIT) values and Calibration Probability (CPO) values, alongside a high R^2 value of 0.8, underscore the model's appropriateness in accurately predicting crime levels and capturing underlying patterns and relationships. Despite some variance between predicted and actual values, as illustrated in Figure 7, the model

performed moderately well. The Mean Absolute Percentage Error (MAPE) value of 29.3% also highlights the extent of variance between predicted and actual crime counts. Although this level of variance was not remarkably low and indicates a need for further refinement of the model, the model's residuals were relatively consistent. Notably, as shown in Figure 8, regions such as Midrand, Sandton, Douglasdale, Roodepoort, Dobsonville, and Moroka were predicted to have high levels of crime, enabling targeted resource allocation and intervention strategies.

In comparison to existing literature on crime prediction methodologies, our implementation of the INLA model offers notable advantages. While previous studies have explored various approaches such as Bayesian networks and regression models (Gorr & Lee 2015; Kang & Kang 2017), they often face limitations in accurately quantifying uncertainties associated with model parameters and handling spatial and temporal dependencies. For instance, conventional models like Poisson regression and Ordinary Least Squares (OLS) regression may struggle with accurately representing complex relationships between predictors and crime counts (Gourieroux et al. 1984; Marra & Radice 2010). Our study demonstrates that INLA addresses these limitations by adopting a probabilistic framework, generating posterior distributions that reflect uncertainties in crime prediction models (as seen in Table A2 and Figure A1 in the Appendix). This facilitates informed decision-making by providing credible intervals for estimated parameters. Moreover, while previous studies have shown that conventional models require subjective selection of basis functions (Louzada et al. 2021), our findings illustrate that INLA successfully automates this process through model comparison and Bayesian model averaging, ensuring a better fit and capturing complex predictor-crime count relationships. Additionally, our results demonstrate that INLA's flexibility allows it to estimate model parameters and uncertainties even in regions with scarce or no reported crime data, overcoming the Poisson model's limitation in extrapolation (Louzada et al. 2021). Furthermore, our study highlights that INLA can incorporate additional spatial or temporal factors, enhancing the representation of crime patterns. While previous studies have demonstrated the efficacy of INLA in crime prediction internationally, our research contributes to filling this research gap by demonstrating its suitability for crime prediction in the distinct nature of South African crime data, providing valuable insights for evidence-based decision-making in crime prevention efforts. Moreover, the generalisability of our results extends beyond CoJ, as the INLA model's flexibility and robustness make it applicable to other regions facing similar challenges in crime prediction.

While our study provides valuable insights into crime prediction using the INLA model, it is important to acknowledge its limitations. Firstly, being a single case study focused on the City of Johannesburg, necessitates further research in diverse contexts. Secondly, biases in the data and potential issues with data quality, such as maintenance, poor record-keeping, or management practices, could introduce inaccuracies or inconsistencies in the predictive model, impacting its reliability and effectiveness. Additionally, the aggregation of data at the police station area level may

overlook nuances and variations within smaller geographical areas, limiting the granularity of our analysis. These limitations may have negatively impacted our model's performance, as seen with the moderate MAPE score, hindering a comprehensive assessment of the INLA model's effectiveness. However, despite these limitations, our study successfully identifies high-crime areas in CoJ and properly ordered regions according to expected levels of crime in the next quarter, fulfilling our primary objective and providing valuable insights for targeted crime prevention efforts.

Reflecting on the study's contributions to crime prediction research, our findings highlight the practical advantages of INLA in addressing the limitations of conventional models and advancing evidence-based decision-making in crime prevention efforts.

7 Conclusions

In this study, INLA was investigated as the model of interest for crime prediction. INLA provided practical benefits, such as automated selection of a basis function and robust uncertainty estimation, making it an appealing choice for this work. Implementing the INLA model, we predicted the prevalence of crime in CoJ, identifying high-crime areas, and the various degrees of crime frequency within different areas. Finally, we showcased how using statistical modelling techniques like INLA may provide reliable crime predictions in data-scarce scenarios, where conventional methods like spatial regression fall short, and can enhance evidence-based decision-making in crime prevention. However, in terms of predictive accuracy, while an R^2 value of 0.8 suggests a strong model fit, the MAPE of 29.3% indicates moderate predictive accuracy. This score highlights that, while the implemented model shows promise and provides various advantages, its performance necessitates further refinement. This performance may be attributed to the limitations of our study, constraining the assessment of the INLA model.

On that note, we highlighted the limitations of this study relying on a single case study and a dataset with limited, aggregated observations. Also, we acknowledged the inherent biases in the crime data stemming from subjective human reporting and potential under-reporting, particularly for theft and vandalism. To mitigate this bias, we calculated percentiles of the posterior distribution to capture uncertainties. However, we recognised that bias may still exist despite our mitigation efforts. In light of this, future research should address these limitations and strive for improvement by broadening the scope to encompass multiple cities, enhancing the accuracy and comprehensiveness of data, and utilising more granular data at the neighbourhood level. Additionally, integrating additional contextual factors such as socioeconomic variables and refining the model with advanced techniques could further improve predictive accuracy.

References

Ahmar, A.S., Adiatma, Aidid, M.K. (2018). Crime modeling using spatial regression approach. *Journal of Physics: Conference Series* (Vol. 954, p. 012013).

Armstrong, J. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Springer US.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225,

Blanes I Vidal, J., & Mastrobuoni, G. (2018). *Police patrols and crime* (Tech. Rep. No. 11393). IZA Discussion Paper.

Boqué, P., Saez, M., Serra, L. (2022). Need to go further: using INLA to discover limits and chances of burglaries' spatiotemporal prediction in heterogeneous environments. *Crime Science*, *11*(1), 1–22,

Borges, J., Ziehr, D., Beigl, M., Cacho, N., Martins, A., Araujo, A., Bezerra, L., Geisler, S. (2018). Time-series features for predictive policing. *2018 IEEE International Smart Cities Conference (ISC2)* (pp. 1–8).

Chainey, S., Tompson, L., Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, *21*, 4–28,

Dubé, J., & Legros, D. (2014). Spatial autocorrelation. In *Spatial Econometrics Using Microdata* (p. 59-91). John Wiley and Sons, Ltd.

Gordon, M.B. (2010). A random walk in the literature on criminality: A partial and critical view on some statistical analyses and modelling approaches. *European Journal of Applied Mathematics*, *21*(4-5), 283–306,

Gorr, W.L., & Lee, Y. (2015). Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, *31*, 25–47,

Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica: Journal of the Econometric Society*, 701–720,

Haining, R. (2001). Spatial sampling. *International Encyclopedia of the Social and Behavioral Sciences*, 14822-14827,

Hu, T., Zhu, X., Duan, L., Guo, W. (2018). Urban crime prediction based on spatio-temporal Bayesian model. *Public Library of Science San Francisco, CA USA*, *13*(10), e0206215,

Institute for Economics and Peace (2023). *Global peace index 2023: Measuring peace in a complex world.* https://www.economicsandpeace.org/wp-content/uploads/2023/09/GPI-2023-Web.pdf.

Jesri, N., Saghafipour, A., Koohpaei, A., Farzinnia, B., Jooshin, M.K., Abolkheirian, S., Sarvi, M. (2021). Mapping and spatial pattern analysis of COVID-19 in central Iran using the Local Indicators of Spatial Association (LISA). *BMC Public Health*, *21*, 1–10,

Kang, H.-W., & Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *Public Library of Science San Francisco, CA USA*, *12*(4), e0176244,

Khan, M., Ali, A., Alharbi, Y. (2022). Predicting and preventing crime: A crime prediction model using San Francisco crime data by classification techniques. *Complexity*, *2022*(1), 4830411,

Liao, R., Wang, X., Li, L., Qin, Z. (2010). A novel serial crime prediction model based on Bayesian learning theory. *2010 International Conference on Machine Learning and Cybernetics* (Vol. 4, pp. 1757–1762).

Louzada, F., Nascimento, D.C.d., Egbon, O.A. (2021). Spatial statistical models: An overview under the Bayesian approach. *Axioms*, *10*(4), 307,

Marra, G., & Radice, R. (2010). Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, *19*(2), 107–125,

Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny.* CRC Press.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S.J., Gelman, A., DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and spatio-temporal epidemiology*, *31*, 100301,

Muchika, I., Ngunyi, A., Mageto, T. (2020). Modeling Burglar Incidents Data Using Generalized and Quasi Poisson Regression Models: A Case Study of Nairobi City County, Kenya. *American Journal of Theoretical and Applied Statistics*, 256-262,

Muff, S., Riebler, A., Held, L., Rue, H., Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *64*(2), 231–252,

Poyton, A., Varziri, M.S., McAuley, K.B., McLellan, P.J., Ramsay, J.O. (2006). Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers & chemical engineering*, *30*(4), 698–708,

R Core Team (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

SAPS (2023). *South African Police Service. Annual Report 2022/23.* https://www.saps.gov.za/about/stratframework/annual_report/2022_2023/Annual-Report-2022-23-final-draft-2023-10-12.pdf.

Tierney, L., & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86,

Urdangarin, A., Goicoa, T., Ugarte, M.D. (2023). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, *36*(2), 333–360,

Vicente, G., Goicoa, T., Ugarte, M.D. (2023). Multivariate Bayesian spatio-temporal P-spline models to analyze crimes against women. *Biostatistics*, *24*(3), 562–584,

Appendix A Crime categories and prediction distributions in CoJ

Table A1: Crime Categories

Murder
Sexual Offences
Attempted murder
Assault with the intent to inflict grievous bodily harm
Common assault
Carjacking
Common robbery
Robbery with aggravating circumstances
Robbery at residential premises
Robbery at non-residential premises
Robbery of cash in transit
Bank robbery
Truck hijacking
Arson
Malicious damage to property
Burglary at non-residential premises
Burglary at residential premises
Theft of motor vehicle and motorcycle
Theft out of or from motor vehicle
All theft not mentioned elsewhere

Table A2: Predictions at 2.5, 50, and 97.5 percentiles of the crime distribution

Regions	Area Code	Posterior mean	0.025 quant	0.975 quant
ALEXANDRA	ALE	126	115	136
BOOYSENS	BOO	105	95	115
BRAMLEY	BRA	72	61	83
BRIXTON	BRI	77	67	87
CLEVELAND	CLE	65	55	75
DIEPKLOOF	DIE	69	59	79
DIEPSLOOT	DIEP	106	95	116
DOBSONVILLE	DOB	130	119	141
DOORNKOP	DOO	38	27	49
DOUGLASDALE	DOU	126	115	136
ELDORADO PARK	ELD	105	94	116
ENNERDALE	ENN	38	28	49
FAIRLAND	FAI	34	23	44
FLORIDA	FLO	65	55	75
HILLBROW	HIL	136	125	147
HONEYDEW	HON	139	128	150
IVORY PARK	IVO	126	114	137
JABULANI	JAB	72	61	82
JEPPE	JEP	92	81	103
JHB CENTRAL	JHB	139	128	149
KLIPTOWN	KLI	66	56	77
LANGLAAGTE	LAN	34	23	44
LENASIA	LEN	64	52	74
LENASIA SOUTH	LENA	36	25	46
LINDEN	LIN	74	64	84
MEADOWLANDS	MEA	92	81	103
MIDRAND	MID	143	132	154
MOFFATVIEW	MOF	59	49	70
MONDEOR	MON	100	90	110
MOROKA	MOR	133	122	144
NALEDI	NAL	40	29	51
NORWOOD	NOR	31	21	42
ORLANDO	ORA	98	87	108
PARKVIEW	PAR	40	30	51
PROTEA	PRO	101	90	111
RABIE RIDGE	RAB	72	61	82
RANDBURG	RAN	92	81	102
ROODEPOORT	ROO	133	123	143
ROSEBANK	ROS	38	27	49
SANDRINGHAM	SAND	34	23	44
SANDTON	SAN	133	123	144
SOPHIA TOWN	SOP	98	87	108
YEOVILLE	YEO	45	34	56

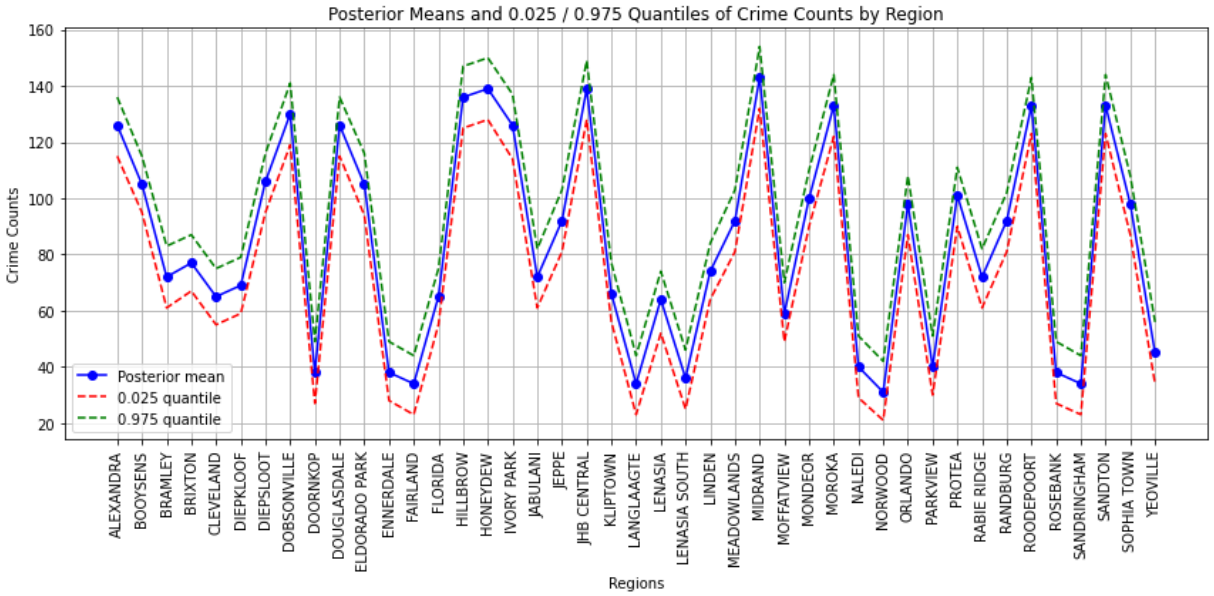


Fig. A1: Graph showing predictions at 2.5, 50, and 97.5 percentiles of the crime distribution.