Network Analysis of City Streets: Forecasting Burglary Risk in Small Areas

Maria Mahfoud[1], Sandjai Bhulai[2], Rob van der Mei[1], Dimitry Erkin[1], and Elenna Dugundji[1]

[1] CWI National Research Institute for Mathematics and Computer Science
[2] Vrije Universiteit Amsterdam, Faculty of Science, Department of Mathematics
Email: M.Mahfoud@cwi.nl, S.Bhulai@vu.nl, R.D.van.der.Mei@cwi.nl, Dimitry.Erkin@gmail.com, E.R.Dugundji@vu.nl

Abstract—**Predicting residential burglary can benefit from understanding human movement patterns within an urban area. Typically, these movements occur along street networks. To take the characteristics of such networks into account, one can use two measures in the analysis: betweenness and closeness. The former measures the popularity of a particular street segment, while the latter measures the average shortest path length from one node to every other node in the network. In this paper, we study the influence of the city street network on residential burglary by including these measures in our analysis. We show that the measures of the street network help in predicting residential burglary exposing that there is a relationship between conceptions in urban design and crime.**

Keywords–predictive analytics; forecasting; street network; betweenness centrality; closeness centrality; residential burglary

I. INTRODUCTION

Residential burglary is a crime with high impact for victims. Substantial academic research has accordingly been dedicated to understanding the process of residential burglary in order to prevent future burglaries [1]. In this attempt, several studies have focused on the role of the urban configuration in shaping crime patterns; this is regarded as one of the fundamental issues in environmental criminology, e.g., [2].

According to [3], environmental criminology is based on three premises. The first premise states that the nature of the immediate environment directly influences criminal behavior, thus a crime is not only reliant on criminogenic individuals, but also on criminogenic elements in the surroundings of a crime. The second premise states that crime is non-randomly distributed in time and space, meaning that crime is always concentrated around opportunities which occur on different moments in a day or week, or different places in a given geographical area. The third premise argues that understanding the criminogenic factors within a targeted environment, and capturing patterns and particular characteristics of that area, can reduce the number of crimes within that area.

Understanding human movement patterns within an urban area is essential for determining crime patterns [4]. These movements occur along a street network consisting of roads and intersections. Throughout the city street network, various places are connected, allowing transportation from one point to the next. Within the network, a street segment can be described as the road, or edge, linking two intersections, or nodes. In their study, [5] found that crime is tightly concentrated around crime hotspots that are located at specific points within the urban area. The urban configuration influences where these hotspots are located, suggesting that it is possible to deal with a large proportion of crime by focusing on relatively small areas. They found that crime hotspots are characterized by being stable over time, and that the hotspots are influenced by social and contextual characteristics of a specific geographical location. To be able to understand and prevent crime, it is important to examine these very small geographic areas, often as small as addresses of street segments, within the urban area. In an analysis of crime at street segment level, [6] reveal that crime trends at specific street segments were responsible for the overall observed trend in the city, emphasizing the need for understanding the development of crime at street segment level.

In urban studies, betweenness is a measure used to determine popularity or usage potential of a particular street segment for the travel movements made by the resident or ambient population through a street network [7], [8]. In criminology, betweenness represents the collective awareness spaces developed by people, including offenders, during the course of their routine activities. This metric provides a means to represent concepts, such as offender awareness, in empirical analysis [9]. Several studies have been conducted to uncover the effects of betweenness on crime. [9] investigated whether street segments that have a higher user potential measured by the network metric betweenness, have a higher risk of burglary. Also included in their research was the geometry of street segments via a measure of their linearity and different social-demographic covariates. They concluded that betweenness is a highly significant covariate when predicting burglaries at street segment level. In another study conducted by [10], a mathematical model of crime was presented that took the street network into account. The results of this study also show an evident effect of the street network.

In this research, we examine for small urban areas (4-digit postal codes: PC4) what the influence of the city street network is on residential burglary by applying betweenness as well as another centrality measure, closeness. These two centrality measures give different indications of the accessibility of an area and we study whether a more accessible area has a higher risk of residential burglary compared to a less accessible area. For comparison, we consider the same areas defined in our previous research [11]. In this earlier study, we predicted residential burglaries within different postal code areas for the district of Amsterdam-West. We extend the model of our earlier research by including the centrality measures closeness and betweenness as explanatory variables. Furthermore, we investigate which of the two centrality measures gives better outcomes, closeness or betweenness.

This paper is organized as follows. Section II describes the dataset and the data analysis. Section III provides the methodological framework of this research. The results of the analysis are discussed in Section IV. In Section V, conclusions and recommendations for further research are presented.

II. DATA

The data used for this research is collected from three different data sources. The first dataset is provided by the Dutch Police and ranges from the first of January 2009 to 30 April 2014. The original dataset includes all recorded incidents of residential burglaries in the city of Amsterdam recorded at a monthly level and grouped into grids of 125 $\times$ 125 meters resulting in 94,224 records. Next to residential burglary, the dataset includes a wide range of covariates. These covariates provide information on the geographic information of the grid such as the number of Educational Institutions (EI) in the grid. In addition to these covariates, the data includes also spatial-temporal indicators of the following crime types: violation, mugging, and robbery. These spatial-temporal indicators measure the number of times a crime type happened within a given grid cell for a given time lag. The second dataset is obtained from the Statistics Netherlands (CBS) and includes various demographic and socio-economic covariates such as the average monthly income. This data is provided on a six alphanumeric postal code level where the first two digits indicate a region and a city, the second two digits indicate a neighborhood and the last two letters indicate a range of house numbers usually a street or a segment of a street. The third dataset is an internal dataset containing different centrality measures calculated on the street network of Amsterdam.

As this research focuses on explaining and predicting residential burglaries at the four-digit postal code level (PC4), the data should be aggregated at this level. Before aggregating the data we perform some pre-processing steps. First, we check the crime records for missing postal codes: if the postal code is missing then all linked data from CBS and the street network will be missing. We observed that 309 of the total 1,812 grid cells had a missing postal code (PC6). Some of these grid cells (34) were subsequently updated manually; other grid cells referred to industrial areas, bodies of water, railroads, grasslands, and highways. As a double check, we also confirmed whether there were residential burglaries in the remaining grid cells with missing postal codes; in our case, there were indeed none. These grid cells were further removed from the dataset and the data were aggregated based on PC4 conditioning on the district as some postal codes (PC4) can cover different police districts. Discrete covariates were aggregated by taking the sum of the covariate on all PC6. For continuous covariates, this was done by taking the average on all PC6. Exploring the data is done in a similar way as discussed in [11], where an extensive data analysis is applied to the crime data and the CBS data. To analyze this data we extend the final set of covariates by the different centrality measures and repeat the same step again. The dataset was assessed for outliers and collinearity. The presence of outliers was graphically assessed by the Cleveland dot plot and analytically by the Local Outlier Factor (LOF) with 10 neighbors and a threshold of 1.3. Results of this analysis show that the training data exhibits a percentage of outliers of 7.6. The majority of these occurred in December and January. Due to the high percentage of outliers in the training set, we decided

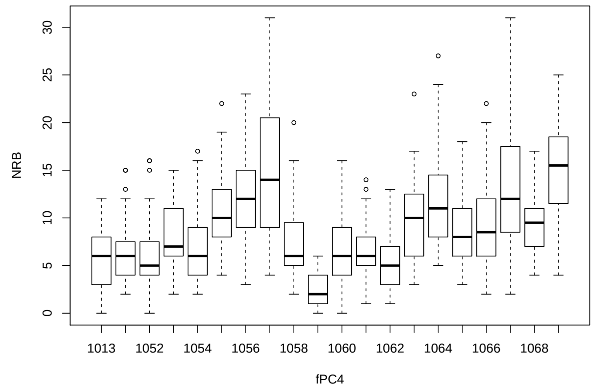

Figure 1. Boxplot of the number of burglaries conditional on the postal code indicating heterogeneity of variance in the number of burglaries within the different postal codes.

to apply the analysis initially without outliers then apply the analysis with the outliers.

The collinearity was assessed by the calculating the variance inflation factor values (VIF) that measures the amount by which the variance of a parameter estimator is increased due to collinearity with other covariates rather than being orthogonal, e.g., [12]. A VIF threshold of 2 is used to assess collinearity [11]. This analysis results in the following set of covariates: the temporal covariate MONTH; the number of educational institutions (sEI), the number of restaurants (sRET), percentage of single-person households (aSH), the number of persons that generate income (sNPI), the total observed mugging incidents in the grid and its direct neighborhood in the last three months (sMuGL3M) and finally, the average monthly income (aAMI).

Furthermore, the relationship between residential burglaries and the categorical covariates was assessed using conditional box plots. Results show a temporal monthly effect and a spatial postal code effect on the burglaries. The effect of the postal codes on the burglaries is illustrated in Figure 1 where a clear difference in the mean and in the variance of the monthly number of burglaries is observed between the different postal codes.

III. METHODOLOGY

A. Centrality measures

Before discussing the centrality measures, we first need to introduce some important concepts of graph theory. A network represented mathematically by a graph is defined as a finite non-empty set V of vertices connected by edges E. A graph is usually written as $G = (V, E)$ where V is the set of vertices and E represents the set of edges where the number of vertices in G is called the order and the number of its edges is called the size. Two vertices u and v are said to be adjacent if there is an edge that links them together. In this case, u and v are also neighbors of each other. If two edges share one vertex then these edges are called adjacent edges. Using this concept of adjacency between all vertices represented in a matrix form results in an adjacency matrix that summarizes all information describing a network.

Another concept for understanding centrality measures is the one of paths and shortest paths. Informally, a path is a way of traveling along edges from vertex u to vertex v without repeating any vertices [13]. Formally, a path P in a graph G is a subgraph of G whose vertices form an ordered sequence, such that every consecutive pair of vertices is connected by an edge. A path P is called an $u-v$ path in G if $P = (u = x_0, x_1, \ldots, x_j = v)$ s.t. $x_0x_1, x_1x_2, \ldots, x_{j-1}x_j$ are all edges of P. The number of edges in a path is called its length. The path $u-v$ with the minimum length is called the shortest path between u and v.

In the context of our analysis, a vertex represents an intersection between streets and an edge is a transport infrastructure supporting movements between the two intersections.

Paths can be considered as the key elements in defining centrality measures. In a transportation network, these centrality measures describe the flow of traffic on each particular edge of the network identifying the most important vertices in it. Some of these centrality measures that we will use in this paper are the closeness (CC) centrality and the betweenness centrality (BC).

Closeness is a very simple centrality measure to calculate. It is a geometric measure where the importance of a vertex depends on how many nodes exist at every distance. Closeness centrality can be defined as the average of the shortest path length from one node to every other node in the network and is given by:

$$\mathrm{CC}(\nu) = \frac{1}{\sum_{d(u,\nu)<\infty} \mathrm{d}(\mathrm{u},\nu)}, \tag{1}$$

where $d(u,\nu)$ is the distance between u and ν. Informally, closeness centrality measures how long it will take to spread information from node ν to all other nodes in the network and it is used to identify influential nodes in the network. The closeness of an edge $u-v$ can be calculated by taking the average closeness values of the nodes u and v.

The betweenness centrality BC is a path-based measure that can be used to identify highly influential nodes in the flow through the network. Given a specific node ν, the intuition behind betweenness is to measure the probability that a random shortest path will pass through ν. Formally, the betweenness of node ν, $\mathrm{BC}(\nu)$ is the percentage of shortest paths that include ν and can be calculated as follows:

$$\mathrm{BC}(\nu) = \sum_{u \neq w \neq \nu \in \mathrm{V}} \frac{\sigma_{u,w}(\nu)}{\sigma_{u,w}}, \tag{2}$$

where $\sigma_{u,w}$ is the total number of shortest paths between node u and w. Moreover, $\sigma_{u,w}(\nu)$ is the total number of shortest paths between node u and w that pass through ν. The betweenness of an edge e can be regarded as the degree to which an edge makes other connections possible and can be calculated in the same way by replacing the node ν by an edge e. An edge with high betweenness value forms an important bridge within the network. Removing this edge will severely hamper the flow of the network as it partitions the network into two large subnetworks.

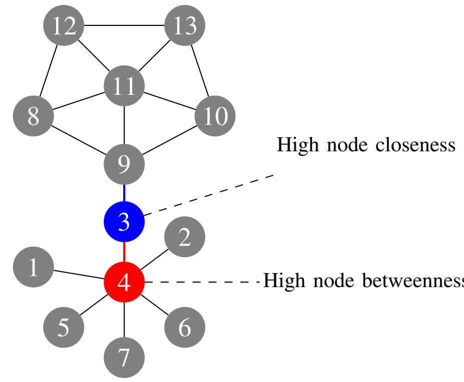

Figure 2. Illustration of high node (edge) betweenness and closeness.

High betweenness or closeness values indicate that a vertex or an edge can reach other vertices or edges, respectively, on relatively short paths. An example of a network is illustrated in Figure 2. In this example, node 3 has the highest closeness and node 4 the highest betweenness. The edge connecting the nodes 3 and 9 has the highest closeness within this network. This edge has also the highest betweenness together with the edge connecting the nodes 3 and 4.

In practice, it is almost impossible to calculate the exact betweenness or closeness scores. To make the calculations feasible, one can set a cut-off distance d and allow only paths that are at distances shorter or equal to d.

B. GAMM including centrality measures

In our paper [11], we used generalized additive mixed-effect models with different structures of the random component and showed that the one-way nested model with postal code as a random intercept has the optimal structure of the random component. Further, we showed that using the population as offset captures the most variation in the data. Moreover, the covariates month and the average monthly income seem to be the most important predictors for the number of burglaries within postal codes. In this paper, the optimal model discussed in [11] will be extended by two different centrality measures as covariates. We assess the effect of these centrality measures on explaining and forecasting the number of burglaries within the postal code. This model is given by:

$$\begin{aligned} &\mathrm{y}_{i,t} \sim \mathrm{Poisson}(\mu_{i,t}), \\ &\mu_{i,t} = \exp(\mathrm{base}_{i,t} + \mathrm{CM}_i + \mathrm{a}_i), \\ &\mathrm{a}_i \sim N(0, \sigma^2_{PC4}), \end{aligned} \tag{3}$$

where a_i is a random intercept for the postal code and $\mathrm{CM_i}$ represents the closeness $\mathrm{CC_i}$ or the betweenness $\mathrm{BC_i}$. The $\mathrm{base_{i,t}}$ is given by:

$$\begin{aligned} \mathrm{base}_{i,t} =& 1 + \mathrm{sEI}_i + \mathrm{sRET}_i + \mathrm{aSH}_i + \mathrm{sNPI}_i + \\ & \mathrm{sMugL3M}_{i,t} + f_1(\mathrm{aAMI}_i) + f_2(\mathrm{Month}_t). \end{aligned} \tag{4}$$

The models were fitted using the Laplace approximate maximum likelihood [14]. This allows comparing the models based on the Akaike Information Criterion (AIC). All analyses were conducted using the `gamm4` package [15].

To assess the predictive performance of the models, the Root Mean Squared Error (RMSE) is calculated for an out-of-sample test. If $\mathrm{y}_{i,t}$ denotes the realization in postal code i and in month t, and $\hat{\mathrm{y}}_{i,t}$ denotes the forecast in the same postal code and in the same month, then the forecast error is given by $\mathrm{e}_{i,t} = \mathrm{y}_{i,t} - \hat{\mathrm{y}}_{i,t}$ and the RMSE is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{e}_{i,t}^2}. \tag{5}$$

C. Space-time model including centrality measures

In our paper [11], we have shown that adding the centrality measure as a covariate to a random intercept model has improved the performance of the model. We also showed that using the closeness as centrality measure with smaller thresholds results in better model performance. The betweenness as a centrality measure leads to better model performance using larger thresholds (larger than 4 minutes). In this paper, we will assess the effect of the centrality measures on burglary risk when modeling the spatial and temporal effects explicitly. Instead of using a random intercept model to account for extra variation within the postal codes, we will model the spatial effect taking into account the spatial autocorrelation.

The main spatial effect ξ_i of area i will be modeled as the sum of a structured effect u_i and an unstructured spatial effect ν_i. The structured spatial effect will be modeled by the mean of a first order intrinsic Gaussian Markov random field [16], [17]. In this specification, the mean of u_i is given by the mean of the adjacent u_i's and the variance of u_i is inversely proportional to the number of neighbors $\mathcal{N}_i$ of area i. The unstructured spatial effect is modeled using exchangeability among the different postal codes. Moreover, the temporal trend of burglary risk is modeled by the mean of a structured and an unstructured component. The temporally structured component is modeled dynamically using a random walk of order 2 and the unstructured component is specified as zero-mean white noise with precision τ_ν.

$$\begin{aligned}
&\mathrm{y}_{i,t} \sim \mathrm{Poisson}(\mu_{i,t}),\\
&\mu_{i,t} = \exp\left(\mathrm{base}_{i,t} + \mathrm{CM}_i + \mathrm{u}_i + \nu_i + \gamma_t + \phi_t\right),\\
&\nu_i \sim N\Big(0, \frac{1}{\tau_\nu}\Big),\\
&\mathrm{u}_i|\{\mathrm{u}_j; j \neq i\}, \tau_\mathrm{u} \sim N\Big(\frac{1}{\mathcal{N}_i}\sum_{j=1}^{n} a_{ij}\mathrm{u}_j, \frac{1}{\mathcal{N}_i\tau_\mathrm{u}}\Big),\\
&\gamma_t|\gamma_{t-1}, \gamma_{t-2} \sim N(2\gamma_{t-1} + \gamma_{t-2}, \sigma^2),\\
&\phi_t \sim N(0, 1/\tau_\phi),
\end{aligned} \tag{6}$$

where a_{ij} is 1 if the area's i and j are neighbors, and 0 otherwise. CM_i represents the closeness CC_i or the betweenness BC_i. The $\mathrm{base}_{i,t}$ is given by:

$$\begin{aligned}
\mathrm{base}_{i,t} =& 1 + \mathrm{sEI}_i + \mathrm{sRET}_i + \mathrm{aSH}_i + \mathrm{sNPI}_i + \\
& \mathrm{sMugL3M}_{i,t} + f_1(\mathrm{aAMI}_i) + f_2(\mathrm{Month}_t).
\end{aligned} \tag{7}$$

The models are fitted using the Integrated Laplace Approximation (INLA) implemented in the R package INLA [18]. The model selection is performed using two selection criterions based on the deviance. First, we use the deviance information criterion (DIC), proposed by Spiegelhalter et al. [19]. The DIC is defined as:

$$\mathrm{DIC} = D(\bar{\theta}) + 2p_D, \tag{8}$$

where $D(\bar{\theta})$ is the deviance using the posterior mean of the parameters, and p_D is the effective number of parameters. As the posterior marginal distributions of some hyperparameters might be highly skewed, especially the precisions, INLA evaluates the DIC at the posterior mode of the hyperparameters. For the latent field, INLA uses the posterior mean [20]. We used also the Watanabe-Akaike information criterion (WAIC) [21]. This criterion is based on the data partition and is closely linked to the Bayesian leave-one-out cross-validation. The WAIC is considered to be an improvement on the DIC criterion [22].

To assess the predictive performance of the models, the Root Mean Squared Error (RMSE) and the Weighted Absolute Percentage Error (WAPE) are calculated using an out-of-sample data set. As before, if $\mathrm{y}_{i,t}$ denotes the realization in postal code i and in month t, and $\hat{\mathrm{y}}_{i,t}$ denotes the forecast in the same postal code and in the same month, then the forecast error is given by $\mathrm{e}_{i,t} = \mathrm{y}_{i,t} - \hat{\mathrm{y}}_{i,t}$. The RMSE is given by Equation (5), and the WAPE is given by Equation (9) defined as follows.

$$\mathrm{WAPE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \mid \mathrm{e}_{i,t} \mid}{\sum_{i=1}^{N}\sum_{t=1}^{T} \mathrm{y}_{i,t}}. \tag{9}$$

IV. Results

In this section, we first present the results of the centrality measures. Then, we will discuss the results of the two models including these centrality measures as covariates.

A. Centrality measures

As discussed in Section III-A, in practice it is computationally very expensive to calculate the exact betweenness and closeness scores. In general, these can be estimated by setting up a buffer zone using a cut-off distance d and calculating these centrality measures by considering only the paths at a shorter length than d. Using historical data, the average speed per street segment was calculated and five different time cut-offs were used. Segments that are reachable within one to five minutes are used to calculate the centrality measures. Note that these averages make sense because the centrality measures are calculated for the whole city and not for each area separately.

The betweenness and the closeness on the street segment level using a cut-off of four minutes are illustrated in Figure 3 and Figure 4, respectively. The corresponding average betweenness and closeness per area are illustrated in Figure 5 and Figure 6, respectively. Figures 3 and 4 show a wide red road running from top to bottom. This road corresponds with the A10, which is the ring road of Amsterdam. Figure 3 also shows that the roads with high betweenness correspond to the main access roads within this district. Figure 4 reveals that the roads within the areas situated on the right-hand side of the A10 have a higher closeness in general. This part of the city was built mainly before the Second World War [23] and has a higher density due to enclosed building blocks creating a

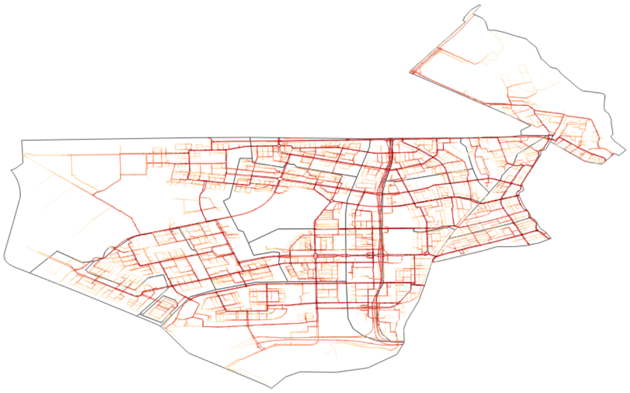

Figure 3. Betweenness of the street segments in Amsterdam West. The betweenness is calculated using the average speed on the street segment and a time threshold of four minutes.

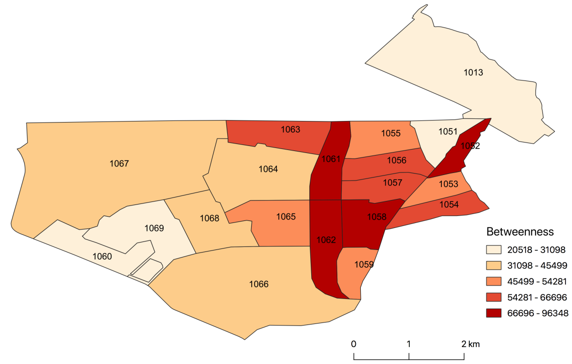

Figure 5. Average betweenness per postal code.

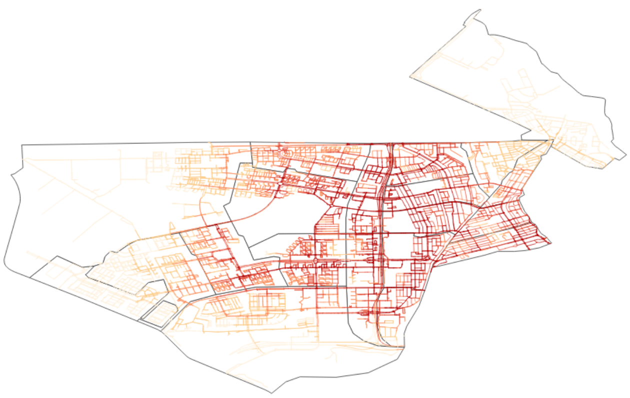

Figure 4. Closeness of the street segments in Amsterdam West. The closeness is calculated using the average speed on the street segment and a time threshold of four minutes.

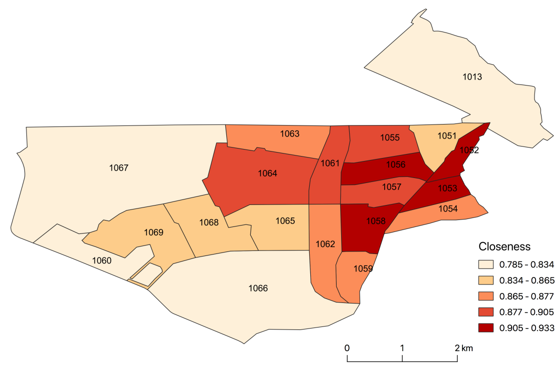

Figure 6. Average closeness per postal code using a threshold of four minutes.

more finely meshed network of roads when compared to the left-hand side of the ring road. This part was built after the Second World War and is characterized by a lower density due to more open building blocks with an emphasis on more green areas and better enclosure of the residential area via main access roads. The blank areas in the district correspond with green areas, such as parks, lakes and agricultural land.

B. *GAMM including centrality measures*

Adding a centrality measure to the GAMM model results in a better prediction based on the RMSE. The RMSE of the GAMM model without centrality measure was about 4.5519 and as can be seen from Table I, extending the model with the betweenness or the closeness results in a generally lower RMSE. It is noteworthy that the closeness leads to better predictions when using lower thresholds (lower or equal $3\ \mathrm{min}$); see Figures 7 and 8. If the threshold is four minutes or higher including the betweenness in the model results in better predictions. This can be explained by the average time an offender might need to flee from the scene of the crime on a residential street to the nearest main access road. In this case, the closeness describes the number of different routes the offender can take during his flight. Within 4 or 5 minutes, the offender can be traveling on the main access road in order to create as much distance as possible from the crime scene.

The results in the area with the postal code 1067 differ from the other areas. Including the closeness and betweenness does not improve the model, the error on the other hand increased. Taking a closer look at this area revealed that this area mainly consists of green areas with few roads. With less alternative routes available, the closeness gives a higher error.

When looking at the other areas, it is possible to say that the

Table I. Root mean squared error (RMSE) values from fitting the GAMM model with closeness and betweenness using different thresholds.

Model	1 min	2 min	3 min	4 min	5 min
GAMM + CC	4.5297	4.5323	4.5366	5.5437	4.5478
GAMM + BC	4.5562	4.5497	4.5405	4.5279	4.5326

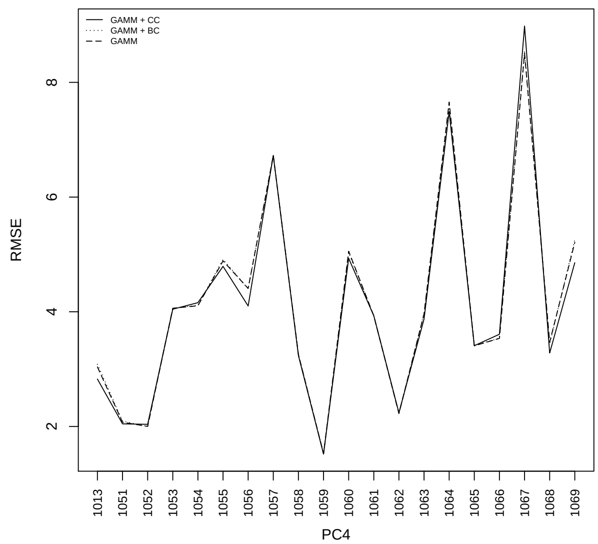

Figure 7. RMSE per PC4 base on an out-of-sample for the GAMM model, the GAMM + CC and the GAMM + BC using a threshold of 1 minute.

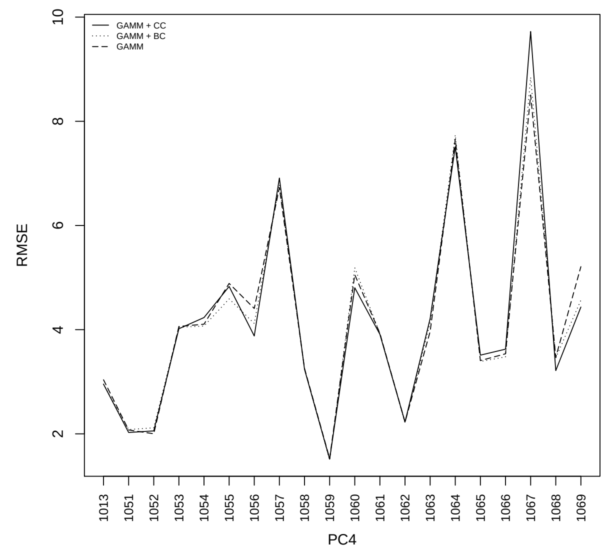

Figure 8. RMSE per PC4 base on an out-of-sample for the GAMM model, the GAMM + CC and the GAMM + BC using a threshold of 4 minutes.

building density influences the effectiveness of the centrality measures on the models. In areas with a lower density, the centrality measures have almost no influence on the outcomes, whereas in the urban areas with a high building density adding the centrality measures to the model improves the outcomes of the model.

Most studies use betweenness as a centrality measure, however, these studies focus on social networks. Given our results, we believe that the closeness is a better centrality measure for modeling crime based on small geographic areas. However, as shown there is a difference in effectiveness of this centrality measure related to the building density of the area.

C. Space-time model including centrality measures

In this section, we will present the results from fitting the space-time models with the betweenness and the closeness using the different thresholds. First, we fit the models including all covariates and compare their DIC and WAIC values. Table II shows that the model with the closeness centrality with a threshold of five minutes provides the lowest DIC and WAIC values. However, the differences remain small. Looking at the estimated posterior mean values and their 95% credible intervals (CI), we can see that the betweenness centrality and the average number of educational institutions are not important as the zero lies within the 95% CI. In contrast, the closeness centrality seems important regardless of the threshold used, see Figure 9.

Based on the DIC and the WAIC values, the best performing model is selected. This model has a closeness measure with a threshold of five minutes and includes all covariates, except the average number of educational institutions and the average number of persons generating income.

The estimated parameters of the best-obtained model on logarithmic scale are presented in Table III. From this table, we can see that the number of retail stores in the postal code, the number of mugging incidents, and the average closeness with a threshold of five minutes have a positive effect on burglaries. The number of households with a single parent has a negative effect on burglaries. To assess the exact effect of these covariates on residential burglaries, we converted the posterior distributions from the logarithmic scale to the original scale of the data. Then we calculated the posterior mean and the 95% credible intervals on the original scale. From Table IV,

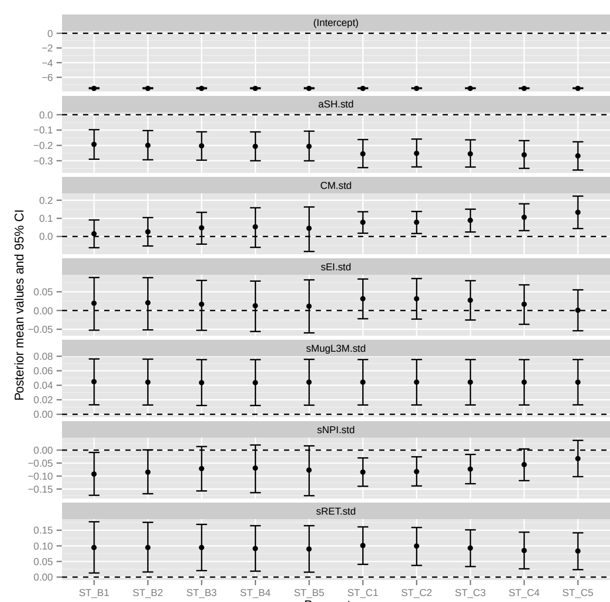

Figure 9. Posterior mean values and 95% credible intervals for all regression parameters obtained using the different models.

Table II. DIC and WAIC values of the models including all covariates.

Criterion	BC1	BC2	BC3	BC4	BC5	CC1	CC2	CC3	CC4	CC5
DIC	3991.795	3991.877	3992.044	3992.049	3991.902	3991.591	3991.725	3991.507	3991.445	3991.434
WAIC	4019.198	4019.266	4019.338	4019.357	4019.263	4018.372	4018.613	4018.255	4018.217	4018.217

Table III. The posterior mean and 95% credible intervals of the fixed effects on logarithmic scale.

Estimate	Mean	Std.dev	0.025 quantile	0.975 quantile
Intercept	-7.486	0.019	-7.525	-7.448
sRET.std	0.086	0.028	0.030	0.142
aSH.std	-0.279	0.044	-0.365	-0.190
sMugL3M.std	0.043	0.016	0.013	0.074
avgCloseness5.std	0.160	0.033	0.094	0.225

Table IV. The posterior mean and 95% credible intervals of the fixed effects on the natural scale.

Estimate	Mean	Std.dev	0.025 quantile	0.975 quantile
Intercept	0.00056085	1.07431e-05	0.00053957	0.000582493
sRET.std	1.09048	0.0303216	1.03077	1.15139
aSH.std	0.757577	0.0331817	0.694364	0.825984
sMugL3M.std	1.04458	0.0161383	1.01311	1.07653
avgCloseness5.std	1.17414	0.038402	1.09919	1.25157

we can conclude that an increase in one unit in the number of retail stores, the number of muggings and in the closeness is associated with an increase of 9.05%, 4.46%, and 17.41%, respectively, in the risk of burglary. Among all covariates, the closeness has the most impact on the risk of residential burglaries. In contrast, the average number of households with a single parent results in a decrease in the risk of burglaries.

After taking account of the covariates, the residual relative risk of each area ($\exp(\xi)$) and their posterior probability of exceeding one ($\Pr(\xi_i > 0 \,|\, \mathrm{y})$) are represented in Figures 10 and 11, respectively. As can be seen from Figure 10, the postal code area 1057 has the highest relative risk of burglaries compared to the whole Amsterdam West. This area also has a higher probability of excess risk on burglaries next to the postal codes 1063 and 1051. These results are in line with our expectations. The postal code area 1057 is a pre-war build neighborhood along the main excess road. These neighborhood houses have many problems such as a higher poverty rate [24] and a higher pollution rate [25]. According to OIS Amsterdam (2017), the adjacent neighborhood inhabits many crime-suspects [24].

The posterior temporal trends are illustrated in Figure 12. This figure represents the structured and the unstructured temporal components modeled dynamically by means of an RW(2) model and an exchangeable Gaussian prior, respectively. As can be seen from this figure, a clear seasonal pattern in burglaries can be observed with a higher risk of burglaries between September and February in general. As expected, peaks are observed for December and January. The same figure reveals that June and July are the months with the lowest risk of burglaries. It is also noteworthy to mention that the second year (2010) clearly has a lower risk of burglaries during the dark months compared to the other years. The temporally unstructured effect, $\exp(\phi_t)$ fluctuates around one.

Finally, we assessed the predictive performance of the models based on out-of-sample data and compared the results to the best obtained GAMM models with the one of the space-time model with a closeness considering a threshold of four

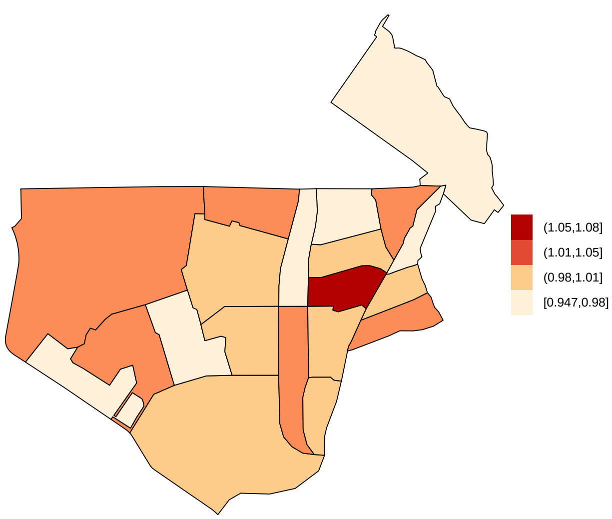

Figure 10. Posterior mean of the residual relative risk for each PC4.

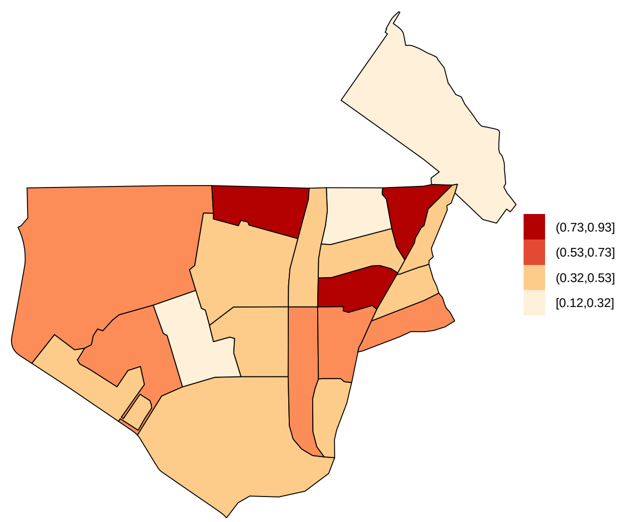

Figure 11. Posterior mean of the residual excess risk for each PC4.

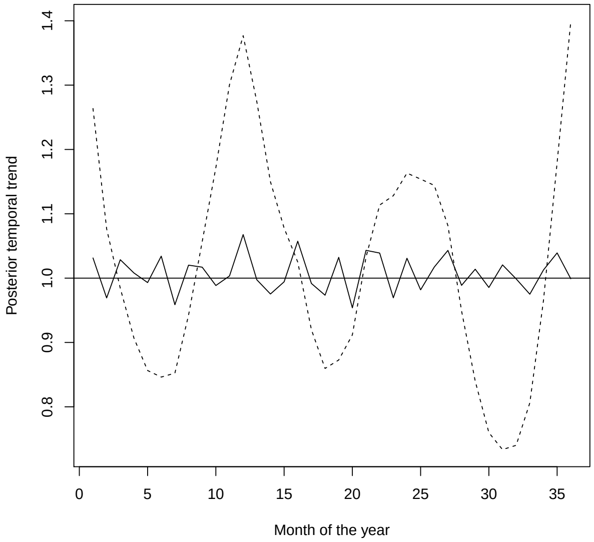

Figure 12. Model with closeness with a threshold of 4 minutes.

minutes. This concerns the GAMM models with the closeness considering a threshold of 1 minute and the GAMM model with the betweenness considering a threshold of 4 minutes. To compare the models we used the RMSE and the WAPE accuracy measures. First, we compared the total performance of the models, then we compared the performance of the models for each postal code separately. The space-time model including the closeness centrality (ST.CC4) clearly results in lower RMSE and WAPE errors compared the GAMM models, see Table V. When we compare the predictive performance of the models for each postal code, we can see that ST.CC4 performs clearly better compared to the other models, especially in PC4 1056, see Figures 13 and 14.

Table V. Predictive performance of the models based on out-of-sample data.

Model	RMSE	WAPE (%)
ST.CC4	3.98	30.47
GAMM.CC1	5.05	36.82
GAMM.BC5	5.35	38.34

V. Conclusion and Future work

During this research, we have tried to determine the influence of accessibility of the street network within small urban areas on residential burglary by applying the centrality measures closeness and betweenness. Given the results in the literature, it is natural to study this problem in the context of GAMM models. We have found that adding the centrality measures as a variable to the GAMM model has improved the performance of this model as can be concluded from the lower RMSE. Furthermore, we have shown that there is a relation between the different conceptions in urban design over time and residential burglary. Our results show that the pre-world War II neighborhoods suffer from more residential burglary than the neighborhoods built after the Second World War.

Rather contrastingly, differences in the performance of the two centrality measures were found when using the GAMM model. Closeness as a centrality measure gives better predictions when taking into consideration a threshold smaller than 4 minutes. If the threshold is 4 minutes or larger, the betweenness gives better predictions. This contradiction disappears when modelling the spatial and temporal effects explicitly. In that case, the model with the closeness centrality with a threshold

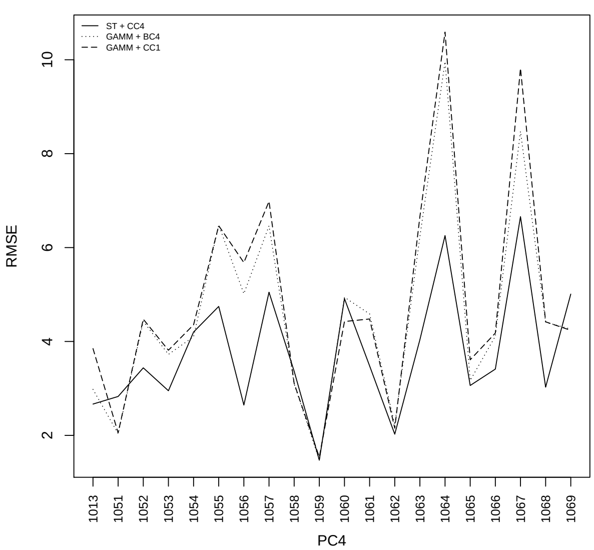

Figure 13. RMSE per PC4 based on out-of-sample data for the ST model including CC5, GAMM + CC1 and the GAMM + BC4.

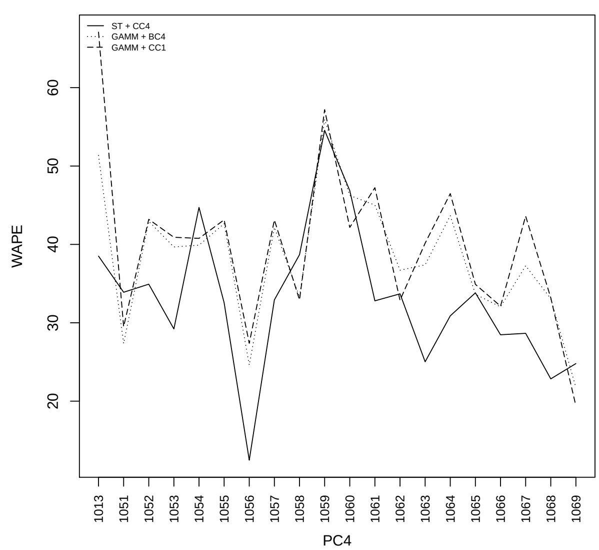

Figure 14. WAPE per PC4 based on out-of-sample data for the ST model including CC5, GAMM + CC1 and the GAMM + BC4.

of five minutes provides the best results, and even beats the best GAMM model.

Our study has shown that there is a relationship between the conceptions in urban design and crime. Neighborhoods built under a certain conception of urban design tend to have a higher risk of residential burglary, which can be explained by how the public space is designed. Further research is necessary to confirm this hypothesis.

REFERENCES

[1] M. Mahfoud, S. Bhulai, and R. van der Mei, “Forecasting burglary risk in small areas via network analysis of city streets,” in Proceedings of the 7th International Conference on Data Analytics. IARIA, 2018, pp. 109–114.

[2] P. J. Brantingham, P. L. Brantingham et al., Environmental criminology. Sage Publications Beverly Hills, CA, 1981.

[3] R. Wortley and M. Townsley, Environmental criminology and crime analysis. Taylor & Francis, 2016, vol. 18.

[4] P. Brantingham and P. Brantingham, “Crime pattern theory,” in Environmental criminology and crime analysis. Willan, 2013, pp. 100–116.

[5] D. Weisburd, E. R. Groff, and S.-M. Yang, The criminology of place: Street segments and our understanding of the crime problem. Oxford University Press, 2012.

[6] D. Weisburd, S. Bushway, C. Lum, and S.-M. Yang, “Trajectories of crime at places: A longitudinal study of street segments in the city of Seattle,” Criminology, vol. 42, no. 2, 2004, pp. 283–322.

[7] L. C. Freeman, “Centrality in social networks conceptual clarification,” Social networks, vol. 1, no. 3, 1978, pp. 215–239.

[8] P. Crucitti, V. Latora, and S. Porta, “Centrality measures in spatial networks of urban streets,” Physical Review E, vol. 73, no. 3, 2006, p. 036125.

[9] T. Davies and S. D. Johnson, “Examining the relationship between road structure and Quantitative Criminology, vol. 31, no. 3, 2015, pp. 481–507.

[10] T. P. Davies and S. R. Bishop, “Modelling patterns of burglary on street networks,” Crime Science, vol. 2, no. 1, 2013, p. 10.

[11] M. Mahfoud, S. Bhulai, R. van der Mei, and D. Kardaras, “Spatio-temporal modeling for residential burglary,” in Proceedings of the 6th International Conference on Data Analytics. IARIA, 2018, pp. 59–64.

[12] D. Liao and R. Valliant, “Variance inflation factors in the analysis of complex survey data,” Survey Methodology, vol. 38, no. 1, 2012, pp. 53–62.

[13] A. Benjamin, G. Chartrand, and P. Zhang, The fascinating world of graph theory. Princeton University Press, 2015.

[14] R. H. Baayen, D. J. Davidson, and D. M. Bates, “Mixed-effects modeling with crossed random effects for subjects and items,” Journal of memory and language, vol. 59, no. 4, 2008, pp. 390–412.

[15] S. Wood and F. Scheipl, gamm4: Generalized additive mixed models using mgcv and lme4, 2014, r package version 0.2-3. [Online]. Available: http://CRAN.R-project.org/package=gamm4

[16] J. Besag, J. York, and A. Mollié, “A bayesian image restoration with two applications in spatial statistics ann inst statist math 43: 1–59,” Find this article online, 1991.

[17] H. Rue and L. Held, Gaussian Markov random fields: theory and applications. CRC press, 2005.

[18] H. Rue, S. Martino, and N. Chopin, “Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations,” Journal of the royal statistical society: Series b (statistical methodology), vol. 71, no. 2, 2009, pp. 319–392.

[19] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, “Bayesian measures of model complexity and fit,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 64, no. 4, 2002, pp. 583–639.

[20] M. Blangiardo and M. Cameletti, Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons, 2015.

[21] S. Watanabe, “Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory,” Journal of Machine Learning Research, vol. 11, no. Dec, 2010, pp. 3571–3594.

[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, Bayesian data analysis. CRC press Boca Raton, FL, 2014, vol. 2.

[23] G. Amsterdam. De groei van amsterdam. [Online]. Available: https://maps.amsterdam.nl/bouwjaar/?LANG=nl (2018)

[24] “Ois amsterdam,” https://www.ois.amsterdam.nl/publicaties/de-staat-van-de-stad-amsterdam/?50159, last access date: 1 March, 2019.

[25] “Wonen in amsterdam 2017: leefbaarheid,” https://www.ois.amsterdam.nl/publicaties/de-staat-van-de-stad-amsterdam/?50159, last access date: 1 March, 2019.