Crime Forecasting in Small City Blocks Using a General Additive Spatio-temporal Model

Maria Mahfoud
CWI,
Stochastics,
Amsterdam, The Netherlands
Email: M.Mahfoud@cwi.nl

Sandjai Bhulai
Vrije Universiteit Amsterdam,
Faculty of Science,
Amsterdam, The Netherlands
Email: s.bhulai@vu.nl

Rob van der Mei
CWI,
Stochastics,
Amsterdam, The Netherlands
Email: R.D.van.der.Mei@cwi.nl

***Abstract*—Spatio-temporal modeling is widely recognized as a promising means for predicting crime patterns. Despite their enormous potential, the available methods are still in their infancy. A lot of research focuses on crime hotspot detection and geographic crime clusters, while a systematic approach to include the temporal component of the underlying crime distributions is still under-researched. In this paper, we gain further insight in predictive crime modeling by including a spatio-temporal interaction component in the prediction of residential burglaries. Based on an extensive dataset, we show that including additive space-time interactions leads to significantly better predictions.**

Keywords–Predictive analytics; forecasting; spatio-temporal modeling; residential burglary.

I. INTRODUCTION

How the police should respond to crime is a constant source of discussion and debate among scholars and practitioners. Over time, new strategies have been developed that use data to influence decision making and direct crime control [1]. This data was first used to indicate the underlying problems within a community by identifying clusters of repeating crime incidents. This was followed by using data to map crime to allow for rapid response to emerging crime problems and hotspots. The most recent development is intelligence-led policing, an objective method for formulating strategic policing priorities by using data analysis and crime intelligence for strategic planning and resource allocation in order to reduce, disrupt and prevent crime. The better integration of the available information systems allows the police to create a picture of the criminal environment and to predict the emerging areas of criminality [2].

Within an intelligence-led framework, proactive policing corresponds with an initial response of the law enforcement agencies to prevent crimes before being committed rather than reacting to criminal acts. Proactive policing requires the ability to predict crime hotspots and concentrations to identify likely targets for police intervention. The identification of these targets is one of the main goals of predictive policing [3].

Although the use of statistical analysis for predicting crimes has been around for decades, the Geographical Information System (GIS) revolution, in the recent years, has led to a surge of analytical techniques to identify likely targets in order to prevent criminal activities. Perry [3] organizes these techniques around six analytic categories: hot spot analysis, regression methods, data mining techniques, near-repeat methods, spatio-temporal analysis and risk terrain analysis. As stated by [4], "the most under-researched area of spatial criminology is that of spatio-temporal crime patterns". The same point has been made by Law et al. [5] who discusses spatio-temporal approaches in past crime research proposing a Bayesian spatio-temporal approach for modeling crime trends. Bernasco and Elffers [6] also address this issue of integrating the spatial and the temporal dimension of crime in order to advance the analysis of crime data. They mentioned that crime varies spatio-temporally illustrating this by an example from [7] on residential burglaries. Especially for residential burglaries, a body of research has shown the repeat and the near-repeat victimization effects [8]–[11]. Therefore, modeling the space-time interactions of residential burglaries are important to capture these effects.

Displaying statistical information on a map allows for conveying information in a format which is ideal for operational decision making. Spatio-temporal information can ideally be understood when displayed on a map, however, there are a number of issues related to the mapping of information in the policing domain. Among these is the use of choropleth maps. As noted by [4], "one particular problem among crime analysis is the incorrect tendency to map real values with choropleth (thematic) maps, resulting in the misleading impression that is often given by larger or unequal areas (Harries, 1999)". Chainey et al. [12] also mention the need of a threshold specification to identify hotspots. In their paper, they indicate also the influence of the parameter setting on the ability to predict future crimes using hotspot maps. The same problem was discussed by [13] who addresses the problem of hotspot identification and the variation of maps that can be obtained using the same data. They state that the choice of a thematic range represents a problem in itself.

An additional problem related to crime mapping is the varying sizes and shapes of geographic administrative boundary areas. Eck et al. [13] propose the use of small uniform grids as a solution to this problem. This results in a high-resolution model. This type of models provides a more realistic forecast in terms of structure and spatial variability [14]. However, it does not necessarily improve the forecast accuracy [15]. Roberts [16] highlights the necessity of evaluating the spatial

and temporal variation in the skill of the model in order to define the scales at which the model forecast should be believed.

This research focuses on residential burglaries and attempts to provide more clarity in predictive crime modeling and mapping by addressing the limitations discussed above. The major aims of this study are to find an accurate probability distribution of residential burglaries taking account of the space-time interactions, and to identify a cut-off value to classify areas as high-risk areas. Wang and Brown [17] model criminal incidents in Charlottesville using a spatio-temporal generalized additive model (ST-GAM) and extend it to a local spatio-temporal generalized additive model (LST-GAM). They applied the ST-GAM to predict the probability distribution of criminal incidents. In the ST-GAM, the temporal information of previous criminal incidents is modeled as a dummy variable indicating the time of the last committed criminal incident. They show that the ST-GAM and the LST-GAM outperform their previous spatial generalized linear model (GLM) and the hot spot model. This research extends the model proposed by [17] by allowing for more complicated space-time interactions.

Inspired by [18], we propose a generalized additive model (GAM) for modeling the probability distribution of residential burglaries in one district of Amsterdam based on regular lattice data (grid boxes of 125 $\times$ 125 meters). The model extends the base model similar to the one discussed in [17] by allowing for additive space-time interactions. We show that the model provides a useful forecast from a radius of 312.5 meters from the centroid of the grid. However, a clear improvement in the forecast accuracy is observed from the first neighborhood (187.5 meters from the centroid of the grid).

The remainder of this paper is organized as follows. Section II describes the used data set and the data analysis. Section III provides the methodological framework underlying this research. Section IV illustrates the results of the analysis. Section V concludes this research.

II. DATA

A. Data description

The data used for this research was provided by the Dutch Police. It contains all recorded incidents of residential burglaries that happened in one district of Amsterdam, with the highest burglary rate, between January 1, 2008 and April 30, 2014. The data was recorded at a monthly level and grouped into grids of 125 $\times$ 125 meters. The data is thus regular lattice data. Only the grids that correspond to urban areas were selected resulting in 1,812 grid locations. In total, there were 115,968 records with a total number of 11,450 incidents.

In addition, each crime incident recorded contained the latitude/longitude coordinates on the grid level, the time of occurrence (month, year) and different covariates that correspond to the demographic factors and the socio-economic factors that are associated with this grid. Next, to these covariates, the Dutch police also use some spatio-temporal indicators that specify when the last incident happened in a specific grid or combination of grids (neighborhood) using different time intervals. These spatio-temporal indicators are crime specific, for example, the number of residential burglaries in a specific grid one month before the reference date. The covariates that correspond to the demographic and the socio-economic factors are location-specific covariates and are constant over time. These covariates count 44 attributes, including population, average values of houses in the postal code area of the corresponding grid, percentage low incomes in the postal code area of the corresponding grid, and so on. Next, to these covariates, we also used some covariates that correspond with the geographic information of the city, such as the distance to the nearest highway access. In total, there were 61 covariates. The description of the discussed covariates is given in Table III.

Table I. Covariates including missing values and the corresponding percentage of the observed missing values.

Covariate	Missing (%)	Covariate	Missing (%)
POP	23	TPH	26
MP	23	ND	31
FP	23	AVH	46
NH	23	NLI	46
AHS	26	NHI	73
NWI	27	NPI	28
SH	26	PB	84
SPH	26	NE	94
MPH	26	AMI	28

B. Data exploration

A first analysis of the recorded incidents shows that only 1.2% of the total records had a higher number of residential burglaries than 1, while 91.61% of the records was equal to 0. For this reason, the occurrence of residential burglaries (binary) was considered as the response variable.

1) Missing values: The first problem encountered using the above-described data was a large number of missing values. The response variable contains no missing values. However, 113,408 of the 115,968 records contain at least one missing value. It is clear that removing every row that contains a missing value is not the best option as it will reduce the sample size by 97.8%.

Further analysis of the missing values shows that all missing values were observed for the location-specific covariates. Moreover, when a covariate contains missing values, at least 23% of the data was missing. Due to a large number of covariates and the high percentage of missing values we decided to remove the corresponding covariates. This concerns 18 of the 44 location-specific covariates. Table I shows these covariates with the corresponding percentage of missing values.

A deeper analysis of the covariates shows that the covariates that correspond to age categories were not complete (they did not sum up to 100%) and at least 25% of the observed values for each variable was equal to zero, which is not likely. For these reasons, these variables were also removed from the data set. Furthermore, the variable TSLI (the number of months since the last incident in the grid) was not always consistent with the corresponding spatio-temporal indicators and based on common sense, this variable is expected to be highly correlated with the other spatio-temporal indicators. For this reason, this variable was also removed from the data set.

2) Near zero-variance covariates: Further analysis of the data shows that many covariates have only some unique values

Table II. Near zero-variance covariates.

Covariate	Covariate	Covariate	Covariate
NA	NS	ACCOM	GI
BANK	SMKT	CS	SCS
NNC	LS	PFS	YC
HOSP	HFE	GH	TO

with low frequencies. These variables, also called near zero-variance variables, can cause numerical problems. Kuhn [19] considered a variable as a near zero-variance variable if two conditions were approved. The first one is that the percentage of unique values should be less than 20%. The second one is that the ratio of the most frequent to the second most frequent value should be greater than 20. The analysis of the near zero-variance covariates in our data set was performed using the `nearZeroVar` function from the caret package [20]. This analysis reveals that 16 covariates have a near zero-variance, which were removed from the data set. These covariates are illustrated in Table II.

The final data exploration was, mainly, performed following the protocol described in [21].

3) Outliers: First, a Cleveland dotplot was drawn for each covariate to identify potential outliers. The plots (see Figures 1-2) show that some covariates have potential outliers indicated by the isolated points. These outliers were replaced by the maximum values observed after removing the outliers from the data set. Moreover, the covariates CB (number of cafes and bars in the grid), REST (number of restaurants in the grid), and SHOP (number of shops in the grid) are highly unbalanced, as illustrated in Figure 1. To avoid problems due to a large number of zeros and to reduce the dimensionality of the data, these covariates were grouped into one covariate called public places (PP). This covariate has 19 unique values but is highly unbalanced. PP was divided into three categories. The first category is when no public places were observed in the grid. The second category is when there are at most five public places in the grid, and the last category is when there are more than five public places. This to distinguish between the grids in terms of crowdedness. Furthermore, EI (the number of educational institutions in the grid) is also highly unbalanced and has only three unique values, this covariate was used as a binary covariate (fPP).

4) Collinearity: Ignoring collinearity increases type II errors and leads to serious problems with forward and backward selection procedures [22]. As we are, among others, interested in the covariates that drive residential burglaries, we should be very careful about collinear covariates. To assess collinearity between the covariates, the variance inflation factor (VIF) was used. The VIF measures the amount by which the variance of a parameter estimator is increased due to collinearity with other covariates rather than being orthogonal [23]. First, the VIF was calculated using all covariates. The covariate with the highest VIF was removed and the VIFs have calculated again. This process was repeated until all VIF values were smaller than two. Note that the use of this threshold is subjective as there is no true VIF threshold. In the literature, different VIF values were suggested. Kennedy [24], among other authors, recommends a threshold of ten. A threshold of five was recommended by [25]. However, as mentioned in [22], the use of a VIF threshold of ten or even five is too high [26]. By using a threshold of two, we aim to be more conservative about collinearity. The VIF analysis shows that L6MN (number of incidents in the direct neighborhood in the sixth month and earlier before the reference time), L6MG (number of incidents in the grid in the sixth month and earlier before the reference time), and ADFS (average distance from the centroids of the grid to the nearest known 10 burglars) are collinear with other covariates and were removed from the data set.

Residential burglaries are known to have the repeat and near-repeat victimization effect where residential burglaries cluster over time and space [8] [27] [10]. Due to this effect, collinearity is expected between the spatio-temporal indicators. To provide more insight into the relationships between these covariates, the principal component analysis (PCA) biplot was used. The left panel of Figure 3 shows higher correlations between the number of incidents observed in the grid and in its direct neighborhood within the same time unit. The spatio-temporal indicators that correspond to the same time unit were aggregated resulting in three covariates TL1M, TL2M, and TL3M where TLxM is the total number of incidents observed in the grid and its direct neighborhood x months before the reference time. A PCA biplot was drawn using these covariates. As can be seen from the right panel of Figure 3, higher collinearity is observed between TL1M, TL2M, and TL3M. Again, to avoid loss of information, these covariates were grouped together into a new covariate, TL3, which is the total observed incidents in the grid and its direct neighborhood in the last three months. To check for outliers in TL3, a Cleveland dotplot was drawn and this plot shows no extreme observations. A PCA biplot was drawn again using TL3, MDFS (distance from the center of the grid to the nearest known burglar) and DTNHA (distance from the center of the grid to the nearest highway access), which shows that MDFS is negatively correlated with TL3 (this plot is not shown here but the same result can be concluded from Figure 3). We decided to use TL3 and leave MDFS out of the analysis.

Furthermore, conditional boxplots were used to assess collinearity between continuous and categorical covariates. This reveals that collinearity between SD and DTNHA exists. The covariate sub-district (SD) also shows some collinearity with TL3. To avoid problems due to collinearity, SD was omitted from the analysis.

The final set of covariates includes eight covariates, namely the space covariates X and Y; the temporal covariates YEAR and MONTH; the categorical covariates public places (fPP) and educational institutions (fEI); the total observed incidents in the grid and its direct neighborhood in the last three months (TL3) and finally, the distance to the nearest highway access (DTNHA).

5) Relationships between the response and the covariates: The relationship between the response variable and the nominal variables was assessed graphically by a design plot (Figure 4). As illustrated in Figure 4, higher mean values of the residential burglaries were observed between October and February, with the highest mean in December. This period is characterized by a short daylight period, while occupancy times of the residents remain the same. Due to the cover of darkness and the absence of the residents, burglars have a lower risk of being spotted. The highest value observed in December can be explained by the Christmas days and New

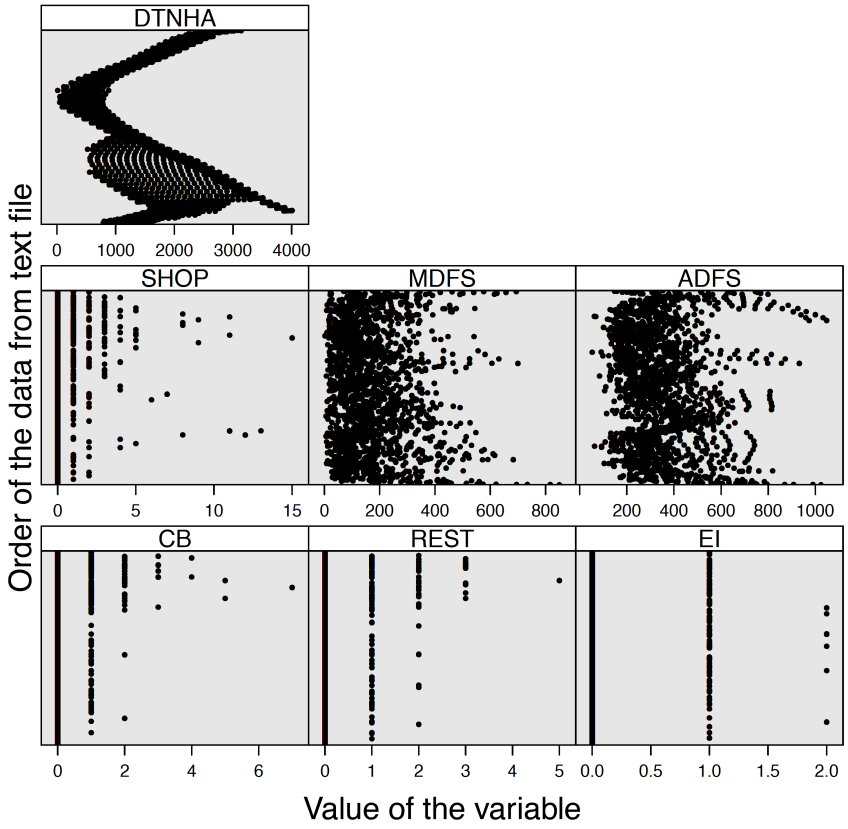


Figure 1. Multi-panel Cleveland dot plot for the location specific covariates. The horizontal axes represent the values of the covariates, and the vertical axes represent the order of the data as imported from the data file. Note the data is sorted on X, Y, YEAR and MONTH, respectively. This figure indicates the existence of some outliers in the most covariates. These are given by the isolated points in the right-hand side of the panels. This figure also shows that the discrete covariates (CB, REST, EI and SHOP) are highly unbalanced with some unique values.

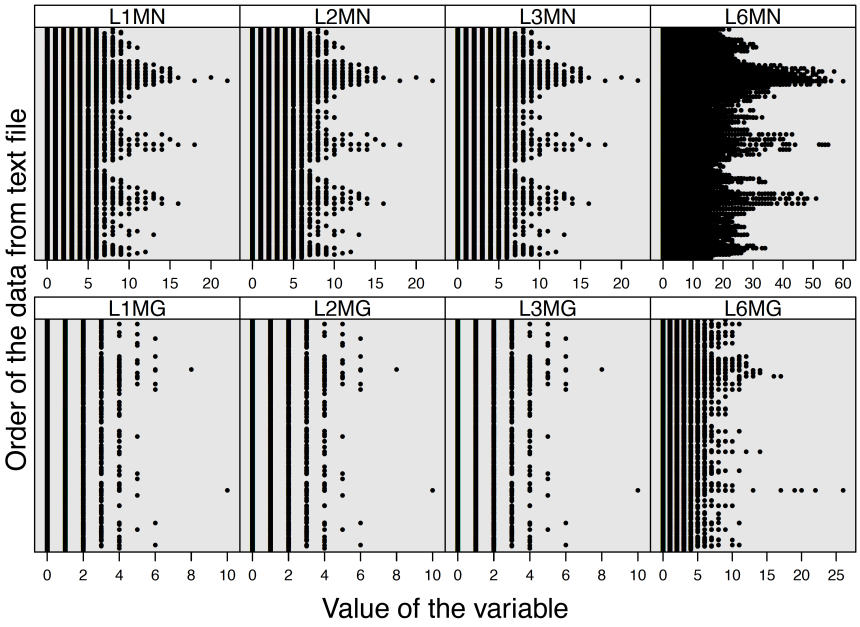
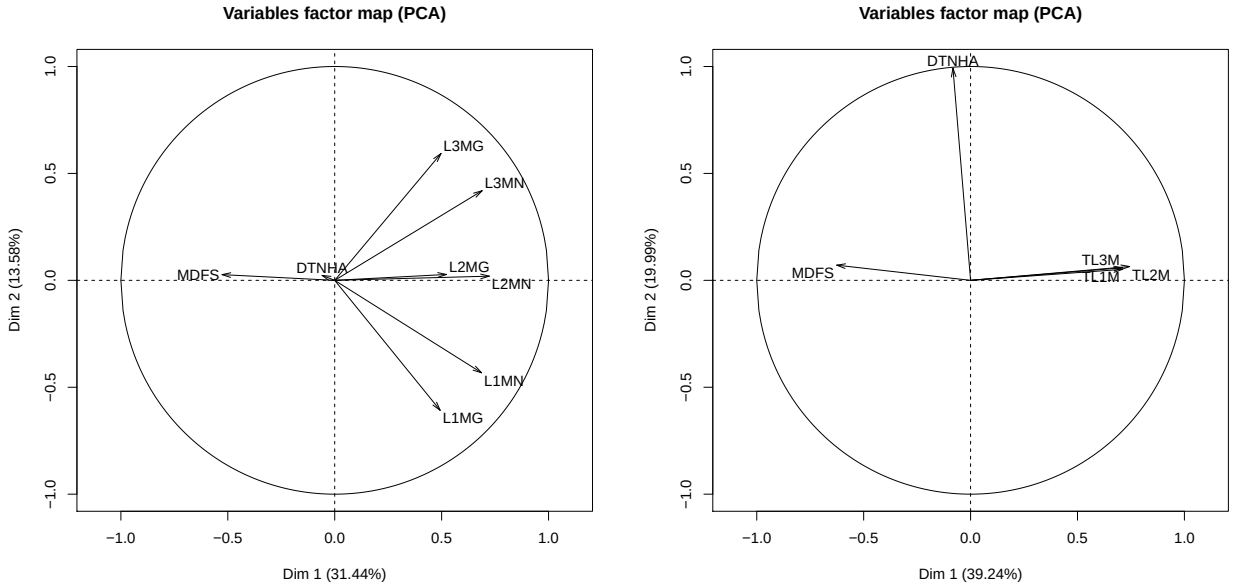


Figure 2. Multi-panel Cleveland dot plot for the spatio-temporal indicators. The horizontal axes represent the values of the covariates, and the vertical axes represent the order of the data as imported from the data file. Note the data is sorted on X, Y, YEAR and MONTH, respectively. This figure indicates the existence of some outliers in these covariates. The panels show roughly the same pattern indicating some collinearity between these covariates.



(a) (b)

Figure 3. PCA biplot of the covariates. The left panel indicates higher correlation between the number of residential burglaries observed in the grid and its direct neighborhood within the same time unit. The right panel shows the PCA biplot after aggregating the spatio-temporal indicators that correspond to the same time-unit. As can be seen from this panel, TL1M, TL2M and TL3M are highly correlated.

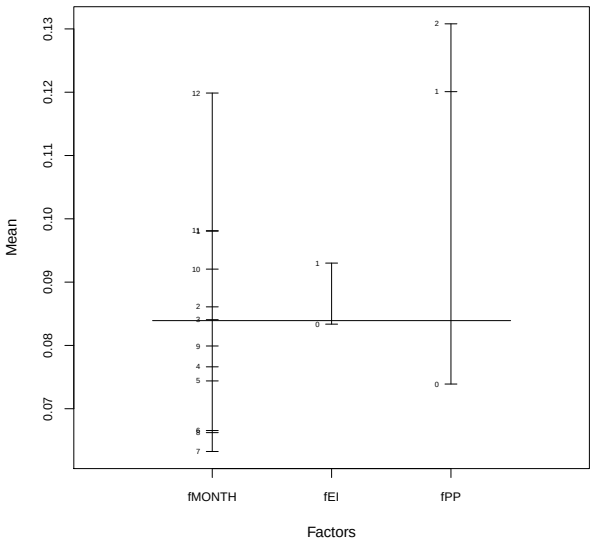


Figure 4. Design plot showing the average incidents per class for each factor variable.

Years Eve that are attractive days for burglars. Furthermore, a higher mean was observed in grids containing educational institutions (fEI) or public places (fPP). Moreover, crowded areas have a higher mean compared to quiet areas.

Finally, histograms of the TL3 and DTNHA for areas with residential burglary were plotted. A deeper analysis on TL3 shows that 93.14% of the incidents has occurred within grids with TL3 higher than zero. For this reason, the histogram of TL3 was drawn considering only TL3 values that are higher than zero. This shows a highly skewed distribution with peaks for TL3 values between two and four. Moreover, the distribution of DTNHA reveals a high peak of residential burglaries for distances between 875 and 1,000 meters.

In the next section, we introduce our generalized additive model (GAM) for modeling the probability distribution of residential burglaries. The model extends a base model by allowing for additive space-time interactions.

III. METHODOLOGY

Given the covariates discussed in Section II, the occurrence of residential burglaries in a certain grid i, and in a certain month t, was modeled using a GAM using the binomial distribution and the logistic link function (see, e.g., [28], [29]). To be more precise, the model is not a GAM with the binomial distribution but rather one with a Bernoulli distribution. The use of the logit link is to ensure that the fitted values are bounded in $(0, 1)$.

The choice of GAM is based on the expected non-linear relationships between the covariates and the response. A non-linear relationship is expected between the response and the distance to the nearest highway access (DTNHA). This can be explained by the two types of burglars identified by [30], the first being the opportunity burglar that prefers to operate within its own neighborhood and the second being the professional burglar who selects its targets based on the highest expected loot and operates mostly in suburban areas and areas that are near highways, because they are unaware of the local situation

and escape routes. A non-linear relationship is also expected for TL3 due to the repeat and near-repeat victimization effects. The covariate MONTH is also expected to have a non-linear effect on the residential burglaries. This is due to the repeat victimization effect and the daylight-darkness effect [31]. For these reasons, smoothers will be used to model these covariates.

We use a GAM model that allows for space-time interactions as follows:

$$\begin{aligned}\text{logit}(\mu_{i,t}) = \text{fEI}_i + \text{fPP}_i + \text{YEAR}_t + f_1(\text{TL3}_{i,t}) + \\ f_2(\text{DTNHA}_{i,t}) + f_3(\text{MONTH}_t) + f_4(\text{X}_i, \text{Y}_i),\end{aligned} \quad (1)$$

where $\mu_{i,t} = \mathbb{E}(y_{i,t})$, $y_{i,t}$ follows a Bernouilli distribution, $i \in \{1, \ldots, 1812\}$, $t \in \{1, \ldots, 60\}$. The functions f_1 and f_2 are one-dimensional smoother functions of the covariates represented by a cubic regression spline (CRS). f_3 is a one-dimensional smoother represented by a cyclic cubic regression spline (CCRS). This is to avoid big jumps between the January and the December value of the smoother [32]. The function f_4 is a two-dimensional isotropic smoother for space represented by thin plate regression splines (TPRS). The TPRS was used for smoothing the spatial co-ordinates because they are measured on the same unit [29].

The model was fitted using the penalized iteratively re-weighted least squares $(\text{P} - \text{IRLS})$, while the optimal amount of smoothing was estimated using the UnBiased Risk Estimator (UBRE) [29]. All analyses were conducted using the `mgcv` package [29] from the R statistical and programming environment [33].

IV. Results

Now that we can generate the probability function of residential burglaries through the GAM model, which cut-off value θ should be used to classify high-risk areas and which spatial scale provides a useful forecast? In practice, the choice of the cut-off value is mostly left to law enforcement agencies who choose a cut-off value based on the available resources and their risk preferences. Some of them choose a cut-off value of 0.8, others select areas based on the top 3% or the top 5% percentiles to classify areas as high-risk areas. However, the use of a hard cut-off value as 0.8 strongly depends on the estimated probabilities. In our case, this will result in a clear under-estimation of risk areas. If one decided to use a fraction of top percentiles, then this should be at least equal to the expected percentage of incidents. Elsewhere, the risk areas will be undoubtedly under-estimated.

Considering our training set, the average incidents (binary) over the five years, ranging between 2008 and 2012, was about 8.3%. This means that on average 151 grids, from the total grids of 1,812, should be considered as risky grids. Using the 97% percentile results in considering only 55 grids as high-risk areas. Doing this, we know apriori that we are under-estimating the risk areas. Some people will argue that the given resources do not allow to cover this high number of grids. In our point of view, from a safety perspective, the grids that should be flagged as high-risk areas should at least match the expected grids with incidents and should be independent of the available resources. After classifying the areas as high-risk areas, smart allocation methods can be used to cover the risk areas using the available resources.

Given the estimated probability distribution, the optimal cut-off (the average) considering the different neighborhoods ($\theta_1 = 0.171$) and the optimal cut-off at the grid level ($\theta_2 = 0.126$) were further used to classify areas as high-risk areas. The reason of using both cut-off values is because the optimal cut-off on grid level was quite different from the optimal cut-offs that correspond to the other neighborhoods.

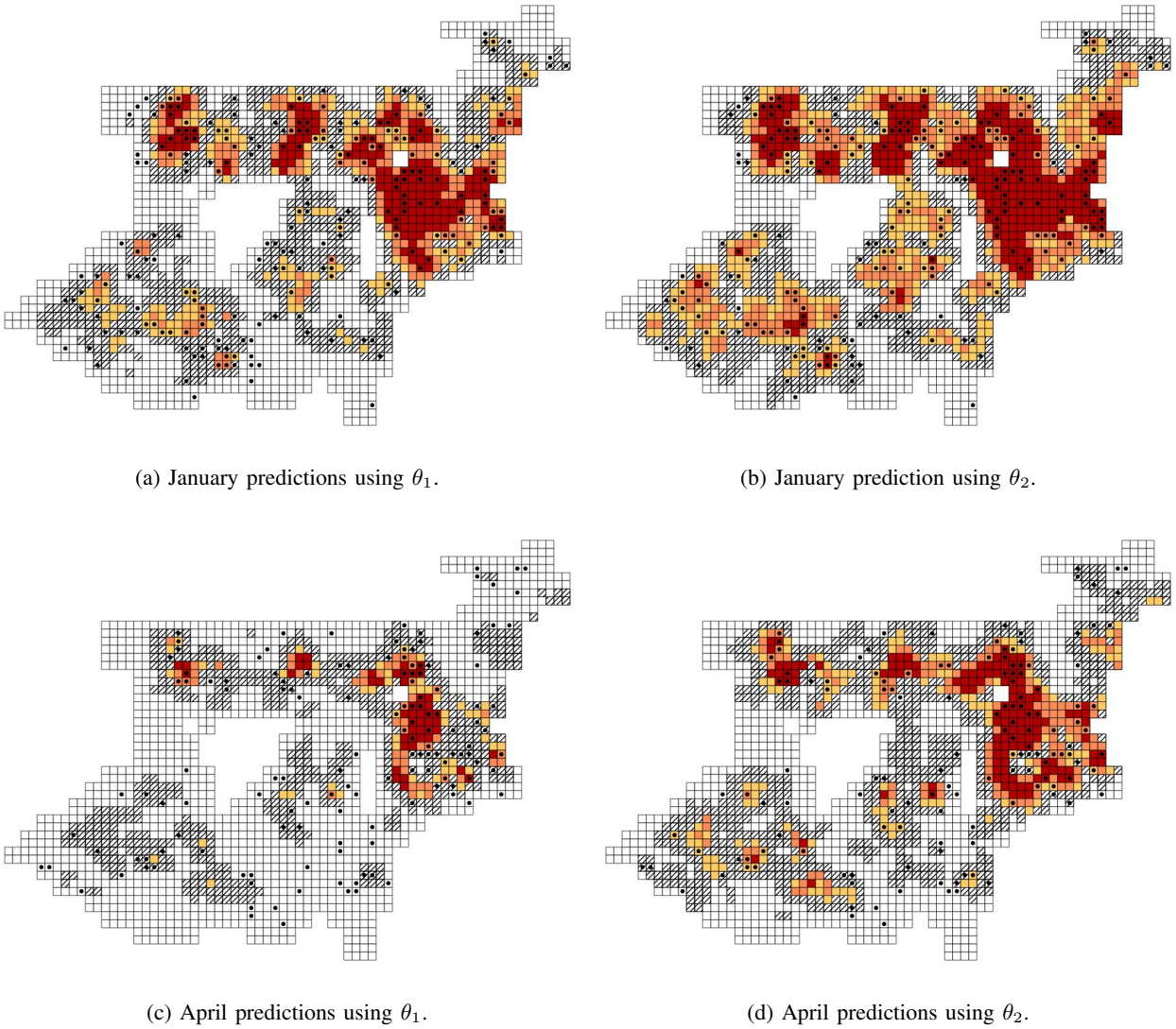
The generated heat maps of January and April are given in Figure 5. From this figure, a clear difference is observed in the number of grids that are flagged as high-risk grids. In fact, more incidents are expected in January compared to April. Therefore, the predicted high-risk area in January is larger compared to the one in April. The heat maps also show that most realizations were located within the high-risk area or within their lower bounds.

In January, more incidents are expected compared to April, this is in agreement with historical data (see Figure 4). The heat maps also show that most realizations are located within the high-risk area or within its lower bound.

V. Conclusion

In this research, we developed a GAM model to predict the probability distribution of residential burglaries. The results show that the covariate TL3, the total incidents in the grid and its neighborhood in the last three months, has a dominant effect in the model. Apparently, this covariate captures a large part of the spatio-temporal effect in residential burglaries. Moreover, a small part of the variation in the data was captured by the model. The low power of the model may be due to the high resolution of the data used.

Finally, θ_1 and θ_2 were used to assess the performance of the model and these cut-offs were compared with the cut-off obtained for the maximum performance. Results show that both values provide similar results to the maximum performance observed, while the cut-offs that correspond to the maximum performance considering the different metrics cover a wide range, which can be difficult to interpret from a decision-making point of view.



(a) January predictions using θ_1.

(b) January prediction using θ_2.

(c) April predictions using θ_1.

(d) April predictions using θ_2.

Figure 5. Heat maps of January and April using θ_1 and θ_2 and including the estimated lower bounds. The heat maps show that almost all incidents are located within the estimated high-risk area or within their lower bounds. It can also be seen that the estimated high-risk area of January is larger than the one of April. The maps obtained using θ_1 show that almost all incidents are located within the high-risk area or within their lower bound. However, the total high-risk area is smaller compared to a high-risk area obtained using θ_2. This result is very appealing for the resource allocation.

REFERENCES

[1] M. Mahfoud, S. Bhulai, and R. v. d. Mei, "Spatio-temporal modeling for residential burglary," in Proceedings of the 6th International Conference on Data Analytics. IARIA, 2017, pp. 59–64.

[2] J. H. Ratcliffe, Intelligence-Led Policing. Willan publishing, 2008.

[3] W. L. Perry, Predictive policing: The role of crime forecasting in law enforcement operations. Rand Corporation, 2013.

[4] J. H. Ratcliffe, "Crime mapping: spatial and temporal challenges," in Handbook of quantitative criminology. Springer, 2010, pp. 5–24.

[5] J. Law, M. Quick, and P. Chan, "Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level," Journal of quantitative criminology, vol. 30, no. 1, 2014, pp. 57–78.

[6] W. Bernasco and H. Elffers, "Statistical analysis of spatial crime data," in Handbook of quantitative criminology. Springer, 2010, pp. 699–724.

[7] J. H. Ratcliffe, "Residential burglars ad urban barriers: a quantitative spatial study of the impact of canberra's unique geography on residential burglary offenders," http://crg.aic.gov.au/reports/ratcliffe.html, 2001, last access date: 31 October, 2017.

[8] W. Bernasco and P. Nieuwbeerta, "How do residential burglars select target areas? a new approach to the analysis of criminal location choice," British Journal of Criminology, vol. 45, no. 3, 2005, pp. 296–315.

[9] S. D. Johnson, W. Bernasco, K. J. Bowers, H. Elffers, J. Ratcliffe, G. Rengert, and M. Townsley, "Space–time patterns of risk: a cross national assessment of residential burglary victimization," Journal of Quantitative Criminology, vol. 23, no. 3, 2007, pp. 201–219.

[10] M. Short, M. D'Orsogna, P. Brantingham, and G. Tita, "Measuring and modeling repeat and near-repeat burglary effects," Journal of Quantitative Criminology, vol. 25, no. 3, 2009, pp. 325–339.

[11] W. Bernasco, S. D. Johnson, and S. Ruiter, "Learning where to offend: Effects of past on future burglary locations," Applied Geography, vol. 60, 2015, pp. 120–129.

[12] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," Security Journal, vol. 21, no. 1, 2008, pp. 4–28.

[13] J. Eck, S. Chainey, J. Cameron, and R. Wilson, "Mapping crime: Understanding hotspots," http://discovery.ucl.ac.uk/11291/, 2005, last access date: 31 October, 2017.

[14] E. E. Ebert, "Neighborhood verification: A strategy for rewarding close forecasts," Weather and Forecasting, vol. 24, no. 6, 2009, pp. 1498–1510.

[15] C. F. Mass, D. Ovens, K. Westrick, and B. A. Colle, "Does increasing horizontal resolution produce more skillful forecasts?" Bulletin of the American Meteorological Society, vol. 83, no. 3, 2002, pp. 407–430.

[16] N. Roberts, "Assessing the spatial and temporal variation in the skill of precipitation forecasts from an nwp model," Meteorological Applications, vol. 15, no. 1, 2008, pp. 163–169.

[17] X. Wang and D. E. Brown, "The spatio-temporal generalized additive model for criminal incidents," in Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on. IEEE, 2011, pp. 42–47.

[18] N. H. Augustin, V. M. Trenkel, S. N. Wood, and P. Lorance, "Space-time modelling of blue ling for fisheries stock management," Environmetrics, vol. 24, no. 2, 2013, pp. 109–119.

[19] M. Kuhn, "Building predictive models in r using the caret package," Journal of Statistical Software, vol. 28, no. 5, 2008, pp. 1–26.

[20] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, and M. Benesty., caret: Classification and Regression Training, 2014, r package version 6.0-37, Last access date: 31 October, 2017. [Online]. Available: http://CRAN.R-project.org/package=caret

[21] A. F. Zuur, E. N. Ieno, and C. S. Elphick, "A protocol for data exploration to avoid common statistical problems," Methods in Ecology and Evolution, vol. 1, no. 1, 2010, pp. 3–14.

[22] A. F. Zuur, E. N. Ieno, and G. M. Smith, Analysing ecological data. Springer New York, 2007, vol. 680.

[23] 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.

[24] P. Kennedy, A guide to econometrics, 6th ed. Willey-Blackwell, 2008.

[25] P. Rogerson, Statistical methods for geography. Sage, 2001.

[26] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis. Wiley, 1992.

[27] S. D. Johnson, "Repeat burglary victimisation: a tale of two theories," Journal of Experimental Criminology, vol. 4, no. 3, 2008, pp. 215–240.

[28] T. J. Hastie and R. J. Tibshirani, Generalized additive models. CRC Press, 1990, vol. 43.

[29] S. Wood, Generalized additive models: an introduction with R. CRC press, 2006.

[30] C. van den Handel, O. Nauta, P. van Soomeren, and P. van Amersfoort, "Hoe doen ze het toch? modus operandi woninginbraak," https://hetccv.nl/onderwerpen/woninginbraak/documenten/hoe-doen-ze-het-toch-modus-operandi-woninginbraak/, 2009, last access date: 31 October, 2017.

[31] T. Coupe and L. Blake, "Daylight and darkness targeting strategies and the risks of being seen at residential burglaries," Criminology, vol. 44, no. 2, 2006, pp. 431–464.

[32] A. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith, Mixed effects models and extensions in ecology with R. Springer, 2009.

[33] R Core Team, R: A Language and Environment for Statistical Computing, 2013, last access date: 31 October, 2017. [Online]. Available: http://www.R-project.org/

Appendix

Table III. List of covariates with a short description.

Covariate	Description
X	X coordinate of grid
Y	Y coordinate of grid
YEAR	Year of reference time
MONTH	Month of reference time
DISTRICT	District
SD	Sub police district
POP	Number of residents in postal code area of the grid
MP	Number of male residents in postal code area of the grid
FP	Number of female residents in postal code area of the grid
NH	Number of households in postal code area of the grid
AHS	Average household size in postal code area of the grid
AC1	Percentage residents between 0 and 14 years old in postal code area of the grid
AC2	Percentage residents between 15 and 24 years old in postal code area of the grid
AC3	Percentage residents between 25 and 44 years old in postal code area of the grid
AC4	Percentage residents between 45 and 64 years old in postal code area of the grid
AC5	Percentage residents between 65 and 74 years old in postal code area of the grid
AC6	Percentage residents 75 years and older in postal code area of the grid
NWI	Percentage non-western immigrants in postal code area of the grid
SH	Percentage of single-person household in postal code area of the grid
SPH	Percentage of single-parent household in postal code area of the grid.
MPH	Percentage of multiple households without children in postal code area of the grid
TPH	Percentage two-parent households in postal code area of the grid
ND	Number of dwellings in postal code area of the grid
AVH	Average value of the houses in postal code area of the grid
NLI	Percentage low income households in postal code area of the grid
NHI	Percentage high income households in postal code area of the grid
NPI	Number of persons that generate income in postal code area of the grid
PB	Percentage of persons that receive social benefits in postal code area of the grid
NE	Percentage of entrepreneurs in postal code area of the grid
AMI	Average monthly income in postal code area of the grid
CB	Number of cafes and bars in the grid
REST	Number of restaurants in the grid
EI	Number of educational institutions in the grid
NA	Number of associations in the grid
NS	Number of snack bars in the grid
ACCOM	Number of hotels in the grid
GI	Number of government institutions in the grid
BANK	Number of banks in the grid
SMKT	Number of supermarkets in the grid
CS	Number of coffee shops in postal code area of the grid
SCS	Number of sex shops, clubs and shows in the grid
LS	Number of liquor stores in the grid
PFS	Number of petrol filling stations in the grid
NNC	Number of nightclubs in the grid
YC	Number of youth centres in the grid
HOSP	Number of hospitals in the grid
HFE	Number of nursing home for the elderly
GH	Number of gambling houses in the grid
TO	Number of tourist offices in the grid
SHOP	Number of shops in the grid
TSLI	Number of months since the last incident in the grid
L1MG	Number of incidents in the grid in the first month before the reference time
L1MN	Number of incidents in the direct neighbourhood of the grid in the first month before the reference time
L2MG	Number of incidents in the grid in the second months before the reference time
L2MN	Number of incidents in the direct neighbourhood of the grid in the second months before the reference time
L3MG	Number of incidents in the grid in the third month before the reference time
L3MN	Number of incidents in the direct neighbourhood of the grid in the third month before the reference time
L6MG	Number of incidents in the grid in the sixth month and earlier before the reference time
L6MN	Number of incidents in the direct neighbourhood in the sixth month and earlier before the reference time
MDFS	Distance (m) from the center of the grid to the nearest known burglar
ADFS	Average distance (m) from the centroids of the grid to the nearest known 10 burglars
DTNHA	Distance (m) from the center of the grid to the nearest highway access