Towards a Burglary Risk Profiler Using Demographic and Spatial Factors

Cristina Kadar*, Grammatiki Zanni, Thijs Vogels, and Irena Pletikosa Cvijikj

Information Management Chair, D-MTEC,
ETH Zurich, Switzerland

{ckadar@ethz.ch,gzanni@student.ethz.ch,
tvogels@student.ethz.ch,ipletikosa@ethz.ch}
http://www.im.ethz.ch/

Abstract. According to modern crime victimization theories, the offender, the victim, and the spatial environment equally affect the likelihood of a crime getting committed, especially in the case of burglaries. With this in mind, we compile an extensive list of potential drivers of burglary by aggregating data from different open data sources, such as census statistics (social, demographic, and economic data), points of interest, and the national road network. Based on the underlying data distribution, we build statistical models that automatically select the risk factors affecting the burglary numbers in the Swiss municipalities and predict the level of future crimes. The gained information is integrated in a crime prevention information system providing its users a view of the current crime exposure in their neighborhood.

Keywords: social issues, open data, risk factors, data mining, crime prevention information systems

1 Introduction

Consistently over time and nations, crime scores as one of the top public concerns [13]. Crime can have different forms and impacts but overall it disturbs the public welfare and reduces the citizens' quality of life [10]. Zooming in on Switzerland for example – once seen as the safest country in Europe, the country has succumbed in the last years to the continents average levels of street violence, burglaries and assault [8]. Within the pie of crime, residential burglary has one of the biggest shares: 70% of all reported crimes represent property crimes, and 14% of these are cases of household burglary [29]. According to the Swiss annual reports on criminal statistics [29], the year 2012 has seen an historical peak of 932 burglaries per year for every 100.000 inhabitants, making Switzerland the top target for burglars in Europe [3,29].

Past research reveals that the number one fact influencing people's decision for buying or renting a real estate is the location of the property [27]. Properties

* Corresponding author.

located in regions where the crime rate is low or exhibits a downwards trend are considered as safer by the potential buyers. While existing information systems (IS) in the public and private sector offer basic visualizations and statistics over the types of different crimes in an area and their evolution over time [22,35], we are currently not aware of any solution providing a deep dive into the local crime factors, and offering its users an accurate and contextualized information of their current crime exposure.

Traditionally, criminology researchers looked at a standard set of demographic factors and built regression models to identify those factors that significantly correlate, either positively or negatively, with the crime rates [21]. But modern victimization theories hypothesize that the environment where the crime happens plays an at least as important role [39]. This is particularly true for burglaries, where the variation in risk of a location being victimized can be explained by factors like the surrounding area, the household and premise characteristics and other aspects of lifestyle affecting the location [20]. This creates an opportunity for non-conventional factors to be integrated in crime prediction models by tapping into open data sources describing our cities, like for example crowd-sourced maps and points of interests (POIs).

In this paper, we investigate a substantial set of heterogeneous factors, we determine if and how they contribute to the amount of residential burglary in the Swiss municipalities, and we make the results available to the public by means of a mobile application. The contributions of the paper are threefold:

1. informed by research in the field of criminology, we leverage various open data sources as relevant burglary predictors: (i) comprehensive census data of the Swiss population, (ii) different categories of POIs in the Swiss communes, and (iii) the national road network;
2. we build statistical models using specific probability distribution assumptions in order to select the model fitting best the data and returning the optimal list of factors affecting the burglary levels;
3. we develop, to the best of our knowledge, the first IS that provides a deep dive into the local crime driving factors, offerings its users a detailed view of the safety in their neighborhood in form of a burglary risk profiler.

The remainder of the paper is structured as follows: we first list existing crime prevention applications, provide a brief overview of background theories from criminological studies, and survey common burglary attractors and detractors in Section 2. We then present our method in Section 3 in terms of the datasets we have leveraged, the pre-processing and data manipulation we have conducted, and the statistical methods we have employed. We present and discuss the results obtained when applying our method to the data in Section 4. In Section 5, we describe the proposed implementation for visualizing the identified risk factors and the predicted crime levels. Finally, we conclude by discussing implications, limitations and future steps in Section 6.

2 Related Work

2.1 Crime Prevention Information Systems

Law Enforcement Solutions To address the social issues of crime, police forces across the world are starting to involve citizens in crime prevention through community policing initiatives [26]. One approach towards citizen empowerment, leading to greater transparency and facilitating group problem-solving strategies, is sharing crime related information with the public in form of crime-tracking information systems, hosted mostly by the police departments themselves. Among the pioneers was the UK Home Office with the publication of monthly crime maps starting from December 2008 [4]. The interest in the Police.co.uk [22] website was so immense that the service crashed on its first day online due to the large number of visitors [12]. Internally, law enforcement agencies have also started integrating crime analytics and prediction components. In this space, one professional solution stands out: PredPol [25], deployed by different police departments in US and used to identify high risk areas in the cities so that patrols can be dispatched accordingly.

Commercial Solutions Apart from the platforms created by the officials, numerous commercial web and mobile applications for crime prevention emerged in the last few years. These solutions fall into two broad categories: (i) crime mapping and (ii) crime sharing IS. Crime mapping IS offer visualization of past crime incidents from trustworthy sources like police departments, sheriff agencies, or news media, in form of individual points or heat maps, with the prime examples of CrimeReports [6], SpotCrime [28], and Trulia [35]. Crime sharing IS facilitate real-time reporting of personal experiences with other users or with the local authorities. The shared information can take the form of (a) personal incidents, like on WikiCrimes [37] and on CrimePush [5], or of (b) tips compiled by users, like on TipSubmit Mobile [34], or by some organization, like on Beat the Burglar Home Survey [1].

2.2 Crime Theories and Burglary Risk Factors

Socio-Demographic Risk Factors of Burglary Early theories trying to explain crime activity in terms of cause or prevention have focused exclusively either on offenders or on victims characteristics [19].

Initially, criminology studies have focused solely on socio-demographic attributes as factors correlating with victimization and have noticed that specific groups of people were facing higher risk of victimization compared with other groups. For example: men, young adults, and African Americans experienced higher risk – while women, old residents and white people had relatively low risk of victimization [14]. Such socio-demographic factors can be included in the *lifestyle-exposure theory* as indicators of lifestyle activities of the victim [14]. In time, various studies falling within the lifestyle theory have found a plethora of factors driving the attractiveness of a property and its risk of being burgled.

Properties with high apparent value [41], properties with owners that have higher education and income levels [19], and properties with few people living inside and which are empty during the day or the night [36] – all have been shown to be more attractive for potential burglars.

On the offender's side, theories like the *social learning theory* and the *theory of anomie and disorganization* deem crime as a product of social disharmony [40]. For instance, population density, unemployment rate and poverty are highlighted as contributing variables to total crime, but also exclusively to burglary [7]. Furthermore, research has found that offenders do not travel far in order to strike and do so based on a crime template derived from their daily routines, traveling paths and overall awareness space [2]. Looking at the demographics of a given spatial unit of analysis, would therefore give insights not only into the victim‘s attributes, but also into the ones of potential criminals.

Socio-demographic factors cannot explain the complete context of crime, as we show below. Yet, they should be included in the models as attributes of both the victims and of the offenders, while not being relied on exclusively.

Spatial Risk Factors of Burglary In an *integrative theory* on the forces affecting crime generation, Miethe and Meier highlight the importance of all three variables in the equation of crime: the offender, the victim, and the situation [19]. The social context works together with the offenders motivation and the victims vulnerability influencing the probability of a criminal event occurring. Even further, the specific environment (seen as the time and place supporting the victimization) is part of the social context where the victim and offender come together and plays a crucial role [39].

Based on a review of the empirical literature, a research brief of the Rutgers Center on Public Security [20] identifies four main aggravating or mitigating spatial correlates specifically of urban residential burglary:

1. proximity to public housing units as operationalization of social disorganization: disadvantaged and disorganized areas have been shown to have high levels of crime due to low levels of informal surveillance, security measures and socioeconomic status, and to high levels of ethnic heterogeneity and residential mobility.
2. proximity to pawn shops: burglars not only want to subtract goods in a quick fashion in order to avoid detection, but they also want to dispose of the stolen items as quickly as possible and convert them to cash.
3. proximity to bus stops: public transportation offers the means for accessing and exiting neighborhoods readily and anonymously.
4. proximity to police stations, fire stations and hospitals: the increased likelihood of authorities being present can be considered a mitigating factor resulting in a decrease of risk in the area.

In the case of suburban residential burglary, burglars may need to secure and use their own means of transportation to move larger products (e.g. televisions). Dwellings that are more permeable, easily accessible and located in big streets with well-established road network are indicated as highly targeted for burglary [9].

Furthermore, prior research has acknowledged that burglars first search for an appropriate offending area and only then scan for a specific target within that area [42]. It is therefore sensible to identify a high risk area on a city or neighborhood level based on the crime correlates creating an environmental setting conducive for residential burglary. Based on this information, it may be then possible at later stages to identify residences within the area that may be at even more risk based on individual characteristics (like employed security measures, lighting, proximity to alleyways etc).

3 Methodology

3.1 Datasets and Pre-Processing

Burglaries Dataset We extracted all burglary incidents with physical damage or after intrusion from an anonymized, aggregated digest of the official police records. This dataset is not public, and was made available to us via a data contract credited in the Acknowledgements section. As the extract was generated early 2014, some cases from 2013 were still not finalized by the police, so complete datasets of all burglaries in Switzerland were available only for years 2012 and older. Below, we will analyze both 2012 and 2011. In 2012, there were 72.996 burglary incidents reported in 2142 of the 2495 official municipalities registered that year. In 2011, there were 67.304 burglaries in 2103 out of 2551 municipalities. Figures 1 and 2 are exemplifying a choropleth map of the burglary density (in incidents/km^2) and a histogram of the burglary counts for 2011 – note that the long tail of the distribution continues until 5256 (city of Zurich) and was cut from the graph. For later analysis, we label every entry as high-crime or low-crime municipality by ordering the datasets on burglary counts and splitting them into two roughly equally-sized subsets.

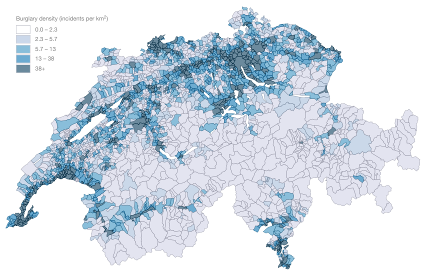

Fig. 1. Switzerland: Burglary density per municipality in 2011.

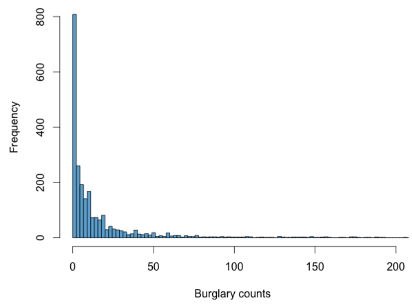

Fig. 2. Burglary counts per municipality in 2011.

Demographics Dataset We acquired a total of 38 socio-demographic and economic factors describing the Swiss population from different governmental sources. Many of these factors were deemed in past studies as significantly influential of residential burglary. Most census data was derived from annual portraits provided by the Swiss Federal Statistical Office [30]: population density, population change (total, through migration, and through natural increase ratios), foreign nationals, age pyramid (young, adult, and old population ratios), spoken languages (German, French, Italian, and Romansh ratios), religion (Protestant, Catholic, and no religion ratios), area usage (settled and used for agriculture/forests/unused ratios), unemployment rate, residents employed in the different economy sectors (primary, secondary, and tertiary sector ratios), residential density (persons per apartment room), main means of transport (public and private ratios), and education level (less than obligatory, secondary, and upper education ratios). Housing (ownership housing, new housing, and vacant housing ratios), gender prevalence (man or woman), and marital status (single, married, and divorced ratios) data was compiled from the Population and Households Statistics [31], while wealth (net income) and social aid information were downloaded from the Interactive Statistical Atlas of Switzerland [32]. All data is already aggregated on the level of municipalities - the lowest administrative unit on which Swiss census data is publicly available - and was collected for years 2011 and 2012. Due to missing values, we remain in the end with 2364 and 2362, respectively, entries per year.

POIs Dataset Widely used in cartography, especially in electronic variants including geographical information systems and GPS navigation software, POIs are specific point locations on a map that might be deemed as useful or interesting for specific activities. They are minimally described by the latitude and longitude or address of the location, type, name, and potentially, description.

We downloaded a POIs dump from the OSM database of Switzerland [23] on 19.02.2015. In total, 136.512 POIs were extracted spanning 167 basic categories, like bars, restaurants, schools, bus stations, or police stations [24]. Because the polygon buffers the Swiss border by 10km and covers Liechtenstein and most of the larger non-Swiss cities along the border, we first filter out all POIs that lay outside the country borders.

We then identify POI duplicates based on the following criteria: (i) being in the same municipality, (ii) having the same type, (iii) having similar names, and (iv) being closely located. For computing the name similarity, we transform the names to lowercase, and use the Smith-Waterman similarity score [33] to find a good alignment of the strings and count the number of identical characters. The computed score is a number between 0 (completely dissimilar names) and 1 (exact matching). In order to obtain a perfect similarity score of 1 for substrings, we normalize this score by dividing it by the string length of the shortest name. To find a good cut-off radius for duplicate search, we look at the proportion of exact name matches between every 2 POIs of the same type within a municipality as a function of their distance: 70% of all POIs pairs within 10m of each other are

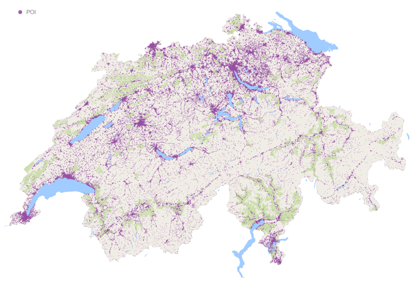

Fig. 3. Switzerland: POIs collected from OSM.

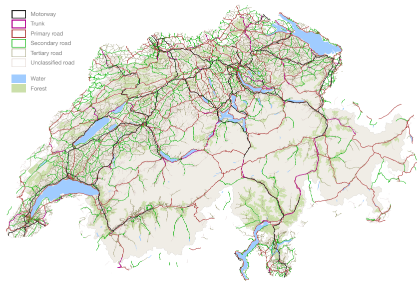

Fig. 4. Switzerland: national road network collected from OSM.

duplicates, 30% at 50m, and 10% at 100m. Based on this, we make the decision to mark as duplicate any pairs of POIs that are in the same municipality, have the same type, have an score of 1 by the metric defined above, and lay within 100m of each other. Only the first point in the alphabetically sorted pair will be used in the further analysis steps.

Figure 3 contains a plot of the 98.585 remaining POIs after the preprocessing steps above. Based on the spatial burglary correlates identified in Section 2.2, we derive a set of four overarching categories: government buildings (including police stations, fire stations, embassies, etc.) restaurants/bars, shops, and public transportation (bus/tram/train stations, airports and so on). For every Swiss municipality, we then count the POIs falling into each of the previous categories; in case no POIs are found, we assume there are none (i.e. zero fill).

Roads Dataset The meta-data of the Swiss road network was downloaded from OSM [23] on 09.05.2015. The extract contains 6 types of roads: motorway, trunk roads, primary road, secondary road, tertiary road, and unclassified roads, and is depicted in Figure 4. As the literature review in Section 2.2 yielded main streets as escape roots and their proximity as an attractor for burglary, for each municipality we compute the motorway and total road density in road km/ha.

3.2 Statistical Approach

In this section, we propose two models for analyzing the effect of the various socio-demographic, economic, and spatial factors described above on the burglary rates of the Swiss municipalities. Our objective is twofold: (i) firstly, we want to build models that accurately predict future crime levels (in terms of expected count numbers and of expected high or low crime exposure), and (ii) equally importantly, understand the relevant risk factors and how they contribute to that.

For the first case, we deal with a response variable $y \in \mathbb{N}$ representing the number of occurrences of a specific event: the burglary counts within a municipality in the course of one year. We face a special case where the distribution of the response variable is discrete and is limited to non-negative integer values. Moreover, the distribution of the data is positively skewed with many observations having low count values – as seen in Figure 2, the data follows a power law distribution. Applying an ordinary linear regression model to these data is problematic. To counteract these problems, Generalized Linear Models (GLMs) of the exponential family need to be employed. The Poisson regression assumes y has a Poisson distribution[1], and that the logarithm of its expected value $E(y) = \mu$ can be modeled as a linear combination of the input variables $\mathbf{x} \in \mathbb{R}^p$: $log(\mu) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{x}_1 + ... + \boldsymbol{\beta}_p\mathbf{x}_p$.

On the side of the explanatory variables $\mathbf{x}$, we deal with many, potentially collinear factors. Including such a high number of variables in the model increases the risk of overfitting. Regularization methods help overcome this risk. Furthermore, we are interested in building a model with good accuracy (i.e. one that has high prediction power), but that still remains easily interpretable (i.e. has high inference power). Towards this goal, we exploit the power of the Least Absolute Shrinkage and Selection Operator (LASSO), a shrinkage method that imposes as regularizer a zero-mean Laplacian prior distribution on model parameters. Thus, LASSO performs automatically variable selection and yields sparse models, i.e. models that involve only a subset of the variables and are much easier to interpret [15].

Below, we deduct the steps necessary for estimating the regression coefficients. Let $\{\mathbf{x}_i, y_i\}_{i=1..n}$ be the input vector of n observation pairs in the training set, and $\{\boldsymbol{\beta}_j\}_{j=0..p}$ the $p+1$ parameters of the model, where $\boldsymbol{\beta}_0$ is the intercept and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_j\}_{j=1..p}$ for ease of notation. The log-likelihood of the training data in the Poisson case is given by:

$$l(\boldsymbol{\beta}_0, \boldsymbol{\beta}|\mathbf{x}, y) = \sum_{i=1}^{n}(y_i(\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\mathbf{x}_i) - e^{\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\mathbf{x}_i})$$

The final objective function to be minimized is the LASSO penalized log-likelihood:

$$\min_{\boldsymbol{\beta}_0, \boldsymbol{\beta}}(-\frac{1}{N}l(\boldsymbol{\beta}_0, \boldsymbol{\beta}|\mathbf{x}, y) + \lambda\|\boldsymbol{\beta}\|_1)$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\boldsymbol{\beta}_j|$ is the l_1 norm of the parameters vector $\boldsymbol{\beta}$. This l_1 penalty has the effect of forcing some of the parameter estimates to be equal to zero when the tuning regularization parameter λ is sufficiently large [15].

For the classification of the crime in risk categories, we choose to implement the LASSO variant of the logistic regression classifier (also known in the literature as the maximum entropy classifier) and deduct below the classification function for the binomial case (i.e. two-class classification). Compared to above, the response variable $g \in G = \{low, high\}$ is a group membership, and not

[1] A discrete random variable Y is said to have a Poisson distribution with parameter $E(Y) = Var(Y) = \mu > 0$, if, for $k = 0, 1, 2, ,$ the probability mass function of Y is given by: $P(Y = k) = \frac{\mu^k e^{-\mu}}{k!}$

a natural value. The model is specified by the following logit transformation: $\log \frac{P(G=high|X=x)}{P(G=low|X=x)} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}^T \mathbf{x}$.

For each observation, it is convenient to code the class membership via a 0/1 response y_i with $y_i = 0$ for $g_i = low$, and $y_i = 1$ for $g_i = high$. Similarly to above, in order to learn the particular values of $\boldsymbol{\beta}$ (in other words, the weights of all predictors for each category), we compute the log-likelihood of the labels on the training data. The final objective function for the LASSO penalized logistic regression will prefer now more sparse models, reducing overfitting and improving performance [15]:

$$\min_{\boldsymbol{\beta}_0,\boldsymbol{\beta}}(-\frac{1}{N}\sum_{i=1}^{n}(y_i(\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\mathbf{x}_i) - \log(1 + e^{\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\mathbf{x}_i})) + \lambda\|\boldsymbol{\beta}\|_1)$$

4 Experimental Results and Discussion

As the LASSO procedure is not invariant to linear transformations of the predictors because of the penalty based on absolute value of $\boldsymbol{\beta}$, we normalize all data before beginning the analysis. In all experiments, we use cross-validation (CV) to compute the parameter estimates for many values of the penalty factor λ, and select the model with the minimal mean error. That means, for a given sequence of potential λs, the program does a k-fold cross-validation, with k = 10, to estimate the prediction error. Each k-times, a model is built based on a fraction $(k-1)/k$ of the data, and the prediction error is computed on the remaining fold. The average error across the k runs is reported and the model with the λ that yields the lowest average error is selected. Furthermore, we apply the obtained model to a dataset of new observations, to test the model's generalization ability on unseen data. Specifically, we will use below the 2011 burglary dataset to build the cross-validated models (i.e. as training and validation sets), and deploy these models to predict the 2012 crime levels (i.e. use the 2012 dataset as test set). For implementation we use the glmnet package, a very efficient R implementation of Lasso and Elastic-Net Regularized GLMs [11].

LASSO Poisson Regression When running the CV experiments, we measure the goodness of fit of the Poisson model by its deviance: $D = 2\sum_{i=1}^{n}(y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i))$ where $\mu_i = e^{\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T\mathbf{x}_i}$ denotes the predicted mean for observation i based on the estimated model parameters. If the model fits well, the observed values y_i will be close to their predicted means μ_i, causing both terms in D to be small, and so the deviance to be small. We show in Figure 5 how the deviance varies with the different values of the regularization parameter λ when estimating a model based on the 2011 values. The optimal $\lambda = 0.7758532$ yields a minimum cross-validated Poisson deviance of 44.58 and selects 15 relevant input factors.

Listed in Table 1 are the corresponding coefficients for all initial $p = 44$ factors. These are to be interpreted as follows: one unit increase in the j^{th} independent variable would lead to a multiplication of the dependent variable y

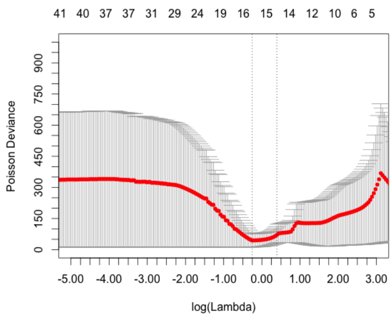

Fig. 5. Poisson regression CV: Poisson deviation as a function of the regularization parameter λ.

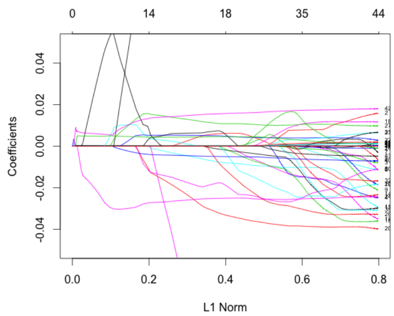

Fig. 6. Poisson regression CV: coefficient values as a function of the L1 norm of λ.

by a factor of $e^{\boldsymbol{\beta}_j}$, given the other variables are held constant in the model. For example: a 1% increase in the unemployment rate of a municipality leads to an increase of the burglary counts (by a multiplication factor of 1.1031), while a 1% increase in the ratio of houses privately owned decreases the burglary counts (the multiplication factor is 0.9750). Furthermore, Figure 6 presents the exact regularization paths of the independent variables, and makes it visible how most get shrunken to 0 eventually.

After the automatic feature selection performed by the LASSO regressor, higher values of the following factors are found to increase the number of burglaries: population density, population change, unemployment rate, population ratio working in the services sector, population ratio using the public transport, population ratio with higher education, number of governmental buildings, and number of stations belonging to the public transport. On the other side, Swiss neighborhoods with the following characteristics exhibit lower burglary counts: higher number of Italian-speaking, Protestant, working in the raw materials industries, with secondary education, and male citizens; higher ratio of privately owned houses; and higher number of restaurants and bars. While most results are consistent with the criminology research presented in Section 2.2, the positive correlation with the number of governmental institutions comes as a surprise. We believe this is due to the bias of the OSM database towards more densely populated areas, where the users were more active and have tagged more POIs.

Moving on to prediction, we apply the learned regressor on new data and estimate the incident numbers for the year 2012. Figure 7 plots the distribution of the actual counts (in blue) vs. the distribution of the predicted counts (in transparent pink). It becomes clear that the empirical data shows more zeroes than would be expected under a regular Poisson distribution, and that we have a good match, but not a perfect one. To formally quantify the accuracy of the estimator we compute the root mean squared error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$: 171.42 in comparison to 34.88 on the training data.

Predictor x_j	Coefficient $\boldsymbol{\beta}_j$	Predictor x_j	Coefficient $\boldsymbol{\beta}_j$	Predictor x_j	Coefficient $\boldsymbol{\beta}_j$
population density	.00009597	settlement surface	.	high education	.00651758
population change	.00517831	empty surface	.	single	.
migration change	.	economic activity	.	married	.
natural change	.	unemployment rate	.09816673	widowed	.
foreigner	.	employed primary	-.02927544	divorced	.
young	.	employed secondary	.	man	-.1120552
adult	.	employed tertiary	.00508251	social aid	.
old	.	residential density	.	net income	.
german	.	ownership housing	-.02531557	government buildings	.00443092
french	.	new housing	.	restaurants and bars	-.004854803
italian	-.00634665	vacant housing	.	shops	.
romansh	.	public transportation	.01165278	transport. stations	.01526131
protestant	-.00013776	private transportation	.	motorway density	.
catholic	.	less oblig. education	.	total roads density	.
no religion	.	secondary education	-.01946257		

Table 1. Poisson regression CV: coefficient values after shrinkage.

LASSO Logistic Regression In a similar fashion, we run the binomial logistic regression CV experiments for year 2011 and obtain an optimal shrinkage parameter $\lambda = 0.0001961$. 31 input variables are preserved in the model, yielding a minimum cross-validated mean classification error of 14.89%, i.e. a mean accuracy of 85.11%.

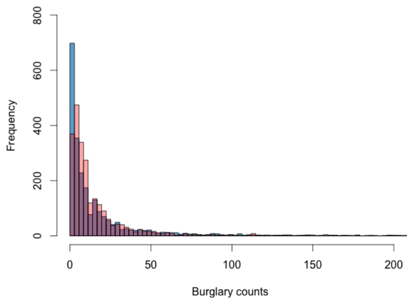

Fig. 7. Poisson regression prediction: histogram of actual vs. predicted burglary counts in the test set.

		Predicted Class	
		Low	High
Actual Class	Low	1020	110
	High	304	928

Table 2. Logistic regression prediction: classification confusion matrix in the test set.

We report in Table 2 the prediction results on the 2012 data in terms of a classification matrix containing the true positives tp, false positives fp, true negatives tn, and false negatives fn counts. The accuracy $Acc = \frac{tp+tn}{tp+tn+fp+fn}$, which represents the percentage of correct predictions made by the model when compared with the actual classifications, is 82.47% – as expected, slightly worse than on the validation set above.

This prediction model performs well on a new target domain, making it a robust choice for deployment in a productive application, as we will show below. The logistic regression classifier is always fitted on the previous year's input factors and burglary data, and, once new input factors are received (e.g. the census releases the new statistics in the beginning of the year) the burglary risk for the current year is computed and shown in the burglary profiler of each municipality.

5 IS Implementation

In order to provide access to the results presented in the previous section to the Swiss citizens, we add a new component to a crime prevention IS we have previously designed and developed. CityWatch [16] intends to increase the safety of its users by raising the awareness about crime levels, as well as by providing concrete crime prevention tips. Figure 8 illustrates the system components, the main view of the crime risk profiler and the visualization of risk factors.

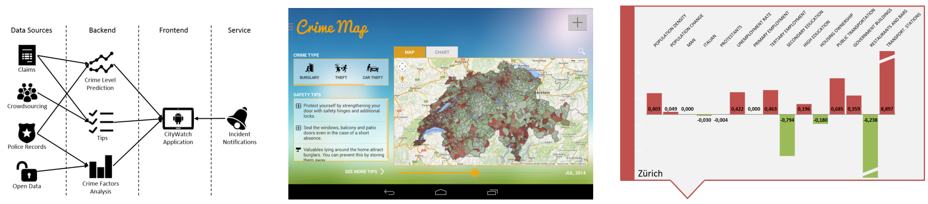

Fig. 8. System components (left), main application view showing aggregated crime map (center) and risk factors (right).

For the crime level information, CityWatch leverages data from two independent sources: (1) property claims data from a large Swiss insurance company, and (2) crowdsourced data. As such, the system provides trustworthy initial information, thus also overcoming the problem of unsustainable content generation and consumption typical for new crowdsourcing applications [17]. At the same time, it provides the possibility to its users to take on an active role in crime prevention, thus contributing to public good [26,38].

For each data source a separate crime map is generated which shows one year of historic data as well as crime prediction for the upcoming year. In order to address the privacy concerns and prevent potential identity disclosure of the victims, the trustworthy map is visualized in an aggregated form thus revealing only the number of incidents on a municipality level but not the exact locations where crime incidents occurred. At the same time, the crowdsourced map reveals the exact incident locations, including details such as date, time, a short description and a photo of the incident, but does not reveal the identity of the reporting user.

In addition to providing the number of incidents, the initial version of the system described in [16] was further extended to serve as a risk profiler by including the specific factors that influence the crime level for a location of interest and their significance based on the analysis presented in the previous section. Moreover, each area is further categorized as safe or unsafe, visualized using a semaphore scheme.

Apart from providing insights into the crime levels and predictors, CityWach counterbalances the potential negative effect that this information might have over individuals by informing its users about possible preventive actions. The preventive tips appear as (a) static tips, which are derived from the recommendations provided by government officials, and (b) dynamic tips, which are generated from the reported incidents and are personalized based on the user profile. Finally, depending on the user‘s preferences, notifications are triggered whenever an incident is reported in user‘s proximity or within any area of interest.

6 Conclusions, Limitations and Future Work

In this paper, we have exploited confidential police criminal records, as well as open data sources, such as census statistics, points of interest, and the national road network. Based on this data, we have built appropriate statistical models in order to identify the relevant risk factors and to predict future burglary rates, and we have shown how the derived information can be integrated in a crime prevention information system. The experimental results mirror some found by other studies in the field, but also reveal other correlates contributing to the empirical literature. Unemployment rates, population employed in the primary sector, houses in private ownership, and number of bus, tram, and train stations are some of the factors we found to influence burglary most strongly. Furthermore, we were able to predict high/low levels of crime exposure with an accuracy of roughly 82% on unseen data.

This work has a broad impact. First, it gives individuals the means to understand the current level of criminality in their neighborhood and the type of risks they are facing. People living in places with higher criminal propensity can increase their awareness and undertake safety precautions. Second, the information could be leveraged by different government institutions. The Swiss police, for instance, could utilize the information by taking effective measures and increasing the patrols in the places predicted with high rates of burglary. Last, enterprises like insurance companies, could run awareness campaigns to inform their clients about the risk they face in order to safeguard their property. In general, such a service brings value to the society at large in form of the presented public platform, but would also find applications in the governmental and private sectors.

Our work is not without limitations and provides several opportunities for further research. We have presented here the results for only 2 consecutive years; to be able to claim that the burglary correlates hold over time, data from a longer time span needs to be analyzed. For more accurate predictions of incident counts, we plan to employ zero-inflated and over-dispersed poisson models in the future. What the IS implementation is concerned: integrating other sources of more dynamic data (like social media or human dynamics) will allow more granular predictions, while looking at other crime types (like pick-pocketing or vehicle theft) will open the way for new use cases.

Acknowledgments. The authors would like to acknowledge the data contract Nr. 140221 (Typ B) with Ref. 650.1-1 from September 2014 for the delivery of the confidential police criminal statistics.

References

1. Beat the Burglar Home Security Survey. Accessed on 24.05.2015: `https://itunes.apple.com/us/app/beat-burglar-home-security/id562269446?mt=8` (2013)
2. Bernasco, W.: A Sentimental Journey to Crime: Effects of Residential History of Crime Location Choice. Criminology, 48(2), pp. 389–416 (2010)
3. Bradley, S.: Who's to blame for the Swiss crime wave? Accessed on 22.05.2015: `http://www.swissinfo.ch/eng/cops-and-robbers_who-s-to-blame-for-the-swiss-crime-wave-/35594462` (2013)
4. Chainey, S., Tompson, L.: Engagement, empowerment and transparency: publishing crime statistics using online crime mapping. Policing, 6(3), pp. 228–239 (2012)
5. CrimePush. Accessed on 22.05.2015: `http://crimepush.com/` (2012)
6. CrimeReports. Accessed on 22.05.2015: `https://www.crimereports.com/` (2015)
7. Cung, B.: Crime Demographics: An Analysis of LAPD Crime Data. Los Angeles, California, United States of America (2013)
8. Dacey, J.: Switzerland loses "safest country" accolade. Retrieved on 22.05.2015 from: `http://www.swissinfo.ch/eng/switzerland-loses--safest-country--accolade/31023508`
9. Davies, P. T., Bishop, R. S.: Modelling patterns of burglary on street. Crime Science a SpringerOpen Journal, pp. 1–14 (2013)
10. Doran, B. J., Burgess, M. B.: Why Is Fear of Crime a Serious Social Problem? In: Putting Fear of Crime on the Map, pp. 9–23. Springer, New York (2012)
11. Friedman J., Hastie T., Tibshirani R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), pp. 1–22 (2010)
12. Garbett, A., Wardman, J. K., Kirman, B., Linehan, C., and Lawson, S.: Anti-social media: communicating risk through open data, crime maps and locative media. In: Proceedings of HCI Korea, pp. 145–152. Hanbit Media, Inc. (2014)
13. Global surveys show environment ranks low on public concerns. Accessed on 26.05.2015: `http://news.uchicago.edu/article/2013/02/25/global-surveys-show-environment-ranks-low-public-concerns` (2013)
14. Hindelang, M., Gottfredson, M., Garofalo, J.: Victims of personal crime: An empirical foundation for a theory of personal victimization. Ballinger Publishing Company, Cambridge, MA. (1978)
15. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York (2009)
16. Kadar, C., Cvijikj, I. P.: CityWatch: the personalized crime prevention assistant. In: Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia, pp. 260–261. ACM (2014)
17. Kumar, V.: Why do Consumers Contribute to Connected Goods? A Dynamic Game of Competition and Cooperation in Social Networks. Social Networks, pp. 1-54 (2009)

18. Mawby, R.I.: Burglary. Willan Publishing, Portland, OR (2001)
19. Miethe, T. D., Meier, R. F.: Crime and its social context: Toward an integrated theory of offenders, victims and situations. In: F. T. Cullen, P. Wilcox, Encyclopedia of criminological theory, pp. 631–635. Sage Publications, Inc. (1994)
20. Moreto, W. D.: Risk Factors of Urban Residential Burglary. RTM Insights, 4. (2010)
21. Osgood, D. W.: Poisson-Based Regression Analysis of Aggregate Crime Rates. Journal of Quantitative Criminology, Vol. 16, No. 1 (2000)
22. London Crime Map. Accessed on 22.05.2015: `http://www.police.uk/metropolitan/00BK17N/crime` (2015)
23. OSM Data for Switzerland. Retrieved from: `http://planet.osm.ch/` (2015)
24. OSM Map Features. Accessed on 22.05.2015: `http://wiki.openstreetmap.org/wiki/Map_Features` (2015)
25. PredPol. Accessed on 22.05.2015: `http://www.predpol.com/` (2015)
26. Ratcliffe, J. H.: Damned if you don't, damned if you do: crime mapping and its implications in the real world. Policing and Society, 12(3), pp. 211–225 (2002)
27. Schwartz, A. E., Susin, S., Voicu, I.: Has Falling Crime Driven New York City's Real Estate Boom? Journal of Housing Research, pp. 101–135 (2003)
28. SpotCrime. Accessed on 22.05.2015: `http://www.spotcrime.com/` (2015)
29. Swiss Federal Statistical Office: Police Crime Statistics, Annual Reports 2011-2014. Retrieved on 22.05.2015 from: `http://www.bfs.admin.ch/bfs/portal/de/index/news/publikationen.html` (2013)
30. Swiss Federal Statistical Office: Portraits of Communes 2011-2012. Retrieved on 10.01.2015 from: `http://www.bfs.admin.ch/bfs/portal/en/index/regionen/02/key.html` (2015)
31. Swiss Federal Statistical Office: STAT-TAB - the FSO's interactive database. Retrieved on 03.03.2015 from: `http://www.bfs.admin.ch/bfs/portal/en/index/infothek/onlinedb/stattab.html` (2015)
32. Swiss Federal Statistical Office: Statatlas Switzerland. Retrieved on 16.02.2015 from: `http://www.bfs.admin.ch/bfs/portal/en/index/regionen/thematische_karten/02.html` (2015)
33. Smith, Temple F., Waterman, Michael S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology, 147, pp. 195–197 (1981)
34. TipSubmit Mobile. Accessed on 24.05.2015: `https://itunes.apple.com/us/app/tipsubmit-mobile/id533472612?mt=8` (2014)
35. Trulia Crime Map. Accessed on 24.05.2015: `http://www.trulia.com/local` (2015)
36. Tseloni, A., Wittebrood, K., Farrell, G., Pease, K.: Burglary Victimization in England and Wales, the United Stages and the Netherlands. British Journal of Criminology, pp. 66–91 (2004)
37. WikiCrimes. Accessed on 22.05.2015: `http://www.wikicrimes.org/` (2015)
38. Wallace, A.: Mapping city crime and the new aesthetic of danger. Journal of Visual Culture, 8(1), pp. 5–24 (2009)
39. Wilcox, P.: Victimization, Theories of. In B. Fisher, S. Lab, Encyclopedia of victimology and crime prevention, pp. 978–986. Sage Publications, Inc. (2010)
40. Williams, F. P., McShane, M. D.: Criminology Theory. Elsevier (1993)
41. Wittebrood, K., Nieuwbeerta, P.: Criminal victimization during one's life course: the effects of previous victimization and patterns of routine activities. Journal of research in crime and delinquency, pp. 91–122 (2000)
42. Wright, R.T., Decker, S.H.: Burglars on the Job: Streetlife and Residential Breakins. Northeastern University Press, Boston, MA. (1994)