Exploring Spatial and Social Factors of Crime: A Case Study of Taipei

Paper ID #62

Abstract. Recognizing the significance of transparency and accessibility of government information, the Taipei Government recently published city-wide crime data to encourage relevant research. In this project, we explore the underlying relationships between crimes and various geographic, demographic and socioeconomic factors. First we collect a total of 25 datasets from the City and other publicly available sources, and select statistically significant features via correlation tests and feature selection techniques. With the selected features, we use machine learning techniques to build a data-driven model that is capable of describing the relationship between high crime rate and the various factors. Our results demonstrate the effectiveness of the proposed methodology by providing insights into interactions between key geographic, demographic and socioeconomic factors and city crime rate. The study shows the top three factors affecting crime rate are educational attainment, marital status, and distance to schools. The result is presented to the Taipei City officials for future government policy decision making.

Keywords: crime factor analysis; geographic information system; demographics; socio-economics; crime hotspots

1 Introduction

Study of crime, criminology, has long been a key area of research spanning across multiple disciplines ranging from behavioral science, sociology, government and education policy planning, to the more recent interdisciplinary data science research. The extreme complexity, and the multifaceted nature of the problem has lead to a recent trend to focus heavily on empirical data analysis approach, with the help of increasing availability of data, and recent advancement in machine learning techniques in data science.

The study of crime often are in two main branches, one focuses on human side of criminal patterns, be it an individual repeated offender, or group/gangs of criminal organizations, with the goal of assisting police investigators in criminal investigations and crime prevention. The other branch focuses on the geographical, spatial, and demographics feature[1] analysis, with the goal of better understanding key factors of why certain areas have higher crime rate (termed *hotspots*). The immediate benefits of this branch of research is more effective law enforcement resource allocation, while the medium and long term goals are to better assist government policy making, which is the focus of our research.

[1] Factor is the term used in Criminology while feature is the technical term used in data science.

2 Related Work

Chen [12] presents a general overview and a framework of using many classic data mining techniques in the first branch, the criminal patterns of human offenders. Wang [21] expands on this with heavy use of machine learning techniques to better the results. On the second branch, there exists many previous works mostly on using GIS system and statistical models and tools like cluster analysis to better understand the correlation between high crime rates with various features [12, 15, 17, 18]. In particular, Ratcliffe [17] works on more accurate prediction of hotspots via better clustering mechanism, with manual inputs of parameters/features from experienced law enforcement agents. Spicer [18] presents a new method for analyzing temporal and spacial crime patterns along major roadways, where where linear spaces are analyzed instead of 2D spaces. This approach allows visualization of temporal variances, crime type comparisons and historical crime trends. Some works focus on analyzing demographic features such as age, education level and divorce rate [6, 14, 19]. Gary [19] in particular, focus on age and explanation of crime, with very detailed yet singleton analysis of age effects on crime behaviors. Bogomolov's work [7] deserves a special mentioning in that it combines the traditional demographic features, such as migrant population, ethnicity, employment and so on, with anonymized and aggregated human behavioral data computed from mobile network activity (called *smartsteps*) to better the prediction accuracy from averaging 50% to 70%.

All of the previous related work exhibits a common theme typical of the modeling and prediction of a complex problem based on incomplete datasets: *the devil is in the detail*, in all aspects of the process, from data collection, cleaning, feature selection/grouping, model building, to repetitive run-through specific model learning and approximation mechanisms. The paper takes the approach of combining spatial and demographic datasets, and uses SVM machine learning techniques to build the most accurate model. This work is summarized as follows: Section 3 details the dataset; Section 4 describes our proposed methodology; Section 5 reports our experimental results and applications; finally, Section 6 concludes our paper.

3 Data Description

Data used in this study are from various government open data repositories and publicly available sources [1–3, 5]. We aggregate these data sources into three datasets, one describing the criminal cases, the other describing demographic characteristics of villages, and the last containing the physical locations of infrastructure and services in Taipei City.

3.1 Criminal Cases Dataset

The criminal cases dataset made available by the City covers three types of crime: burglary, car theft and bike theft[2], covering the time span from January 2015 to April 2016, with a total of 746 household theft, 132 car theft, and 452 bike theft records. Each record contains the crime ID, time (within 3 hours), date, and street address.

[2] Taipei is a relatively safe city with extremely low violent crime rate.

3.2 Village Profiling Dataset

Villages, under districts, are the fourth administrative subdivision of Taiwan, and there are, in total, 456 villages in Taipei. This dataset describes demographic characteristics of the village and includes the following data fields: boundary, area, population, sex ratio[3], average age, ageing index[4], average income, number of households, average electricity usage per household, average number of people per household, ratios of education levels[5] and ratios of marital status[6]. The datasets from various sources are matched with their village ID.

3.3 Point of interest Dataset

This dataset includes the Point of Interest of infrastructure and services with potential impact on crime factors, including parks, banks, bus stops, Metro stations, street surveillance cameras, public restrooms, convenience stores, factories/warehouses, public parking lots, street-side parking spaces, police stations and city street lamps. Note that the location of schools include the location of kindergarten, elementary, junior and senior high school. Because these datasets come from a variety of sources with different coordinate systems, as part of the data cleaning process, we use Google Maps Geocoding API for converting street addresses to WGS84-coordinates, and tools from the Information Science Institute from Academia Sinica [4] for converting TWD97-coordinates to WGS84-coordinates.

4 Methodology

In this section, we apply the grip thematic mapping technique [8,10] to combine the different layers of spatial data, including criminal cases dataset, village profiling dataset and point of interest dataset to predefined 500 meters by 500 meters cells. We extract various features from the above datasets and exploit feature selection and ranking methods to identify high impact features. We then build a data-driven model using the selected features to predict the probability of high crime rate of each individual cell. The remaining section details this experimental process.

4.1 Granularity Definition and Data Preprocessing

Each field in our datasets, such as location of police stations and marital status of villages, can be thought of as individual spatial layers. However, these data are not valuable or informative unless we define a suitable referencing system to combine and relate the

[3] Sex ratio is defined as the ratio of males to females in a population.

[4] Ageing Index is calculated as the number of persons 60 years old or over per hundred persons under age 15.

[5] Ratios of education levels include the six percentages: The percentage of population illiterate, with elementary school, junior high, senior high, undergraduate and graduate education.

[6] Ratios of marital status include the four percentages: The percentage of population single, married, divorced or widowed.

different spatial layers. In order to address this, we apply the grid thematic technique and draw a grid over Taipei that creates 500 meters by 500 meters cells as visualized in Figure 1. The reasons are two fold:

1. Manually defining the grid ensures that each of our boundaries enclose the same area. Because the variance in the area of villages is too large, it is more difficult to perform proximity analysis that can be generalized to each village.
2. This approach allow us to control the fidelity of our combined dataset. When the cells are too small, the village datasets obtained via village-wide statistics are too coarse to describe the characteristic of individual cells; when the cells are too large, the level of detail of cells deteriorates.

Furthermore, multiple studies have shown the effectiveness of this type of spatial analysis [7, 10]. As the result, we defined 1081 cells in Taipei City, and, for each of the defined cells, we perform the following preprocessing steps.

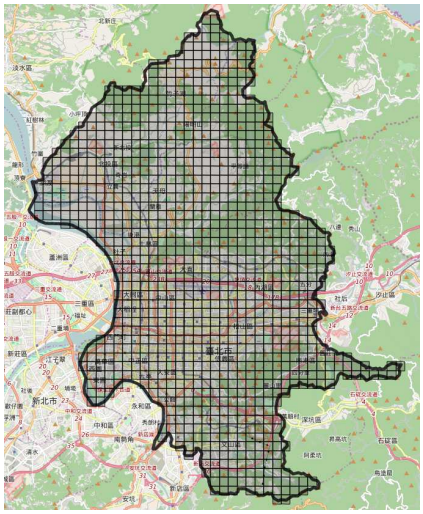

Fig. 1. Visualization of every 500*500 cell in the Taipei City.

Mapping crime locations to cells. We georeference crime locations to cells by assigning it to its closest cell, and in order to avoid spatial autocorrelation [15], we calculated *crime rate* by normalizing the number of crimes by the number of buildings within each cell with crime occurrences. In order to build a binary classification model, we split the transformed crime dataset into two classes with respect to its median: low crime rate (class ”0”) when the index is less than or equal to its median, and high crime rate(class ”1”), when the index is higher than its median. The empirical distribution function of the index is shown in Figure 2 and the median is 0.019. In addition, because the dataset is split with respect to its median, the distribution of classes is balanced.

Mapping village profiles to cells. We georeference the village dataset to each cell by calculating the percentage of the cell in nearby villages to create a weighted sum for the village dataset. In order to decrease computational cost, we created 100 points evenly spaced within the cell and sorted them into nearby villages to estimate the percentage of each cell in nearby villages. Furthermore, in order to avoid collinearity between education level ratios, we combine the six ratios into one metric: the percentage of population with undergraduate or graduate level of education. We also generate another set of features by transforming the non-ratio features into log-scale As the result, there are 19 features generated from the village profiling dataset.

Mapping point of interest dataset to cells. In order to relate the point of interest dataset to cells, we use proximity analysis where we associate spatial features through their physical proximity. For each datasets, there are two metrics: one is to count the number of objects within the boundary of the cell, and the other is to calculate the distance between the center of the cell and its nearest object. Furthermore, we generate another metrics by transforming the distance features into log-scale. Thus, there are 36 point of interest features created for a cell.

4.2 Feature Selection

After we combine our datasets, we normalize each feature by taking standard scores to increase the convergence speed of our algorithms [9], and ease our analysis later on. In order to perform out-of-sample validation to our model, we randomly separate one-fifth of our dataset from our sample. The result is a training set with 80% of our data and testing set with 20% of our data. We only use the training set in the following feature selection process.

In our preliminary data analysis, we perform a Pearson correlation analysis to identify correlation between each feature, as well as with the response variable. Because of high correlation between several of our features and in order to avoid overfitting, we decide that feature reduction is necessary. As the result, we use a feature ranking generated from random forest, as well as Pearsons correlation test to perform an initial feature selection process. For each point of interest, we keep only one metric out of the three generated from 4.1 and for each characteristic of a village, we keep only one metric out of the two generated from 4.1 if a log-scaled feature is created. We also remove any features that has high correlation to other features and low correlation to our response variable. As the result, we reduce our feature space from 55 features to 29 features.

4.3 Model Construction

We then train a variety of binary classifiers with the training data using 5-fold cross validation, including logistic regression, naive bayesian, decision tree, random forest and support vector machine with different parameters and kernels if applicable. For each classifier, we determine its optimal set of features by using a SVM-based recursive feature elimination (RFE) algorithm [11]. In this algorithm, we rank the features by

importance using the linear SVM feature ranking and recursively eliminated the feature with the lowest significance while calculating the performance of the model. In order to increase robustness, the linear SVM feature ranking is the average of our result from each iteration of our 5-fold cross validation process and the top 15 features are illustrated in Table 1. The RFE process allows us to graph the AUC performance against the number of features used to identify the optimal set of features each model should use. The SVM-based RFE graph for each binary classifier is illustrated in Figure 3.

Table 1. Features ranked by SVM-RFE

Rank	Features	Min.	1^{st}-quantile	Median	3^{rd}-quantile	Max.
1	log.village.average.peopleperhousehold	0.694	0.925	0.986	1.028	1.131
2	log.village.average.population	7.050	8.466	8.684	8.911	9.254
3	log.nearestdistance.kindergarden	2.159	4.873	5.316	5.833	7.667
4	village.ratio.undergraduate	0.275	0.463	0.526	0.584	0.708
5	log.nearestdistance.anyschool	2.159	4.803	5.192	5.713	7.551
6	log.nearestdistance.highschool	3.454	6.033	6.526	6.888	8.274
7	village.ratio.married	0.373	0.452	0.469	0.484	0.643
8	log.nearestdistance.bank	1.902	4.920	5.600	6.272	8.407
9	log.nearestdistance.mrt	2.568	5.807	6.318	6.868	8.659
10	log.nearestdistance.policestation	3.685	5.708	6.235	6.573	7.763
11	log.nearestdistance.publicrestroom	1.160	4.460	5.011	5.443	7.437
12	village.ratio.divorced	0.035	0.053	0.061	0.071	0.132
13	village.ratio.single	0.340	0.421	0.436	0.449	0.583
14	log.nearestdistance.juniorschool	3.148	6.029	6.437	6.841	8.050
15	village.average.income.per.person	6.559	6.907	7.063	7.291	8.527

Note that calculate (i) distances in unit of meters, and (ii) population in unit of thousands.

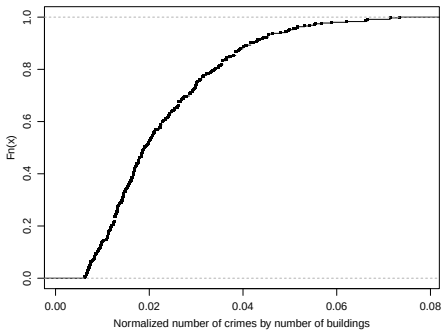

Fig. 2. The empirical cumulative distribution function of the crime rate

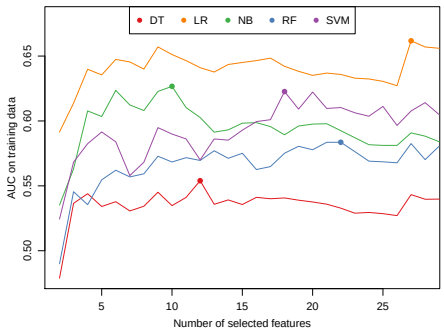

Fig. 3. AUC curve of each model with respect to number of selected features

5 Experiment Results

In this section we compare our experimental results between the five models mentioned in Section 4.3 . We train the models on four-fifth of all cell samples in the training set and test using the one-fifth out of sample testing set. For each model, we use its accuracy, F1 score, and the area under the ROC curve (AUC) to evaluate the performance. We then analyze the implications of our model.

5.1 Model Comparison

Table 2 concludes the performance metrics of the five models we explore in our experiment. The SVM model outperforms the other models in all metrics, achieving 73.4% accuracy in predicting high crime rate versus low crime rate. Figure 5 depicts the spatial visualization of predicted results. Hence, we adopt this SVM model to analyze how the demographic, geographic and socioeconomic features influence the high crime rate probability of a cell.

In order to further analyze the interaction between high crime rate probability and the various features, we apply the Platt Scaling method, which trains the parameters of an additional sigmoid function that maps the output scores of our SVM classification model into posterior class probabilities [13, 16]. This transforms our binary classification model into a regression model that predicts the probability of a cell having high crime rate, or its *high crime rate probability*.

Table 2. Model performance comparison

Classification model	# Used features	Accuracy	F1 score	AUC
Decision tree (DT)	12	0.671	0.629	0.668
Logistic regression (LR)	27	0.646	0.650	0.660
Naive Bayesian (NB)	10	0.671	0.594	0.657
Random forest (RF)	21	0.696	0.684	0.705
Support vector machine (SVM)	18	**0.734**	**0.747**	**0.756**

5.2 Quantification of Feature Impact

First of all, we conduct exploratory study on the correlation between each feature and high crime rate probability of cells and there are several critical implications. As Figure 6 indicates, the highly negative correlation depicts that the cells where more people have received higher education have lower high crime rate probability. In Figure 6 we also observe that the lower average income would also accompany with higher high crime rate probabilities, and according to the study [20], this may be the result from the unaffordable financial cost of having burglary protection devices like window grill. To further quantify the significance of each feature on high crime rate probability, we create a cell that contains the features median value, as listed in Table 1, for each feature. We then, for each feature, perform the following procedure:

1. Vary the features value from its median to 1^{st}-quantile and to 3^{rd}-quantile while constraining the rest of the features at their median.
2. Evaluate these two newly created cells in our model to calculate their respective high crime rate probability.

The importance of a feature is therefore quantified as the difference between the high crime rate probability when varying it from its 1^{st}-quantile to its 3^{rd}-quantile. The calculated significance of each feature is illustrated in Figure 7 and we make several observations: (i) increase in educational attainment best reduces high crime rate probability (ii) the further the distance away from more populous residential areas, marked by schools and parks, the higher the high crime rate probability, and (iii) decrease in stability of marital status, marked by being divorced or widowed, correlates to increased high crime rate probability.

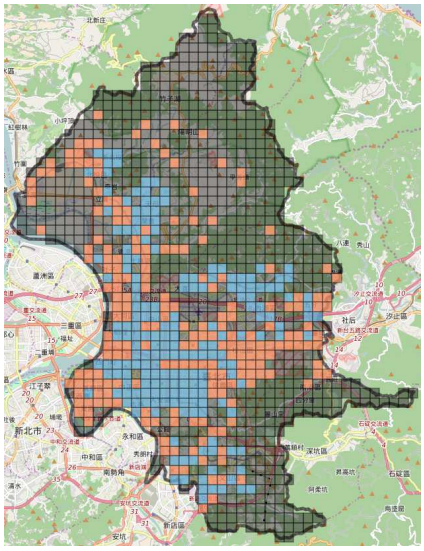

Fig. 4. Predicted by the SVM model

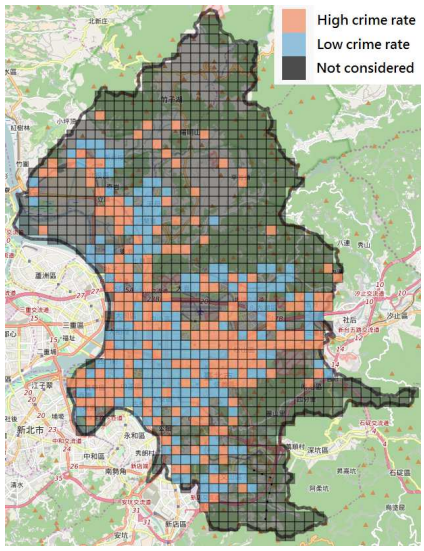

Fig. 5. Ground truth

5.3 Applications

Our experimental results provide insights on how high crime rate probabilities of cells are influenced by their geographic, demographic and socioeconomic features. We can use this knowledge to reveal the underlying causes of crime occurrence and advise government officials accordingly. For instance, given a region whose extrinsic factors are potentially risky like high ratio of divorced, this becomes somewhere the Taipei government can pay more attention upon. Furthermore, our predictive model can also be used on those cells without crime occurrences yet to identify cells with similar characteristics as known places with high crime rate.

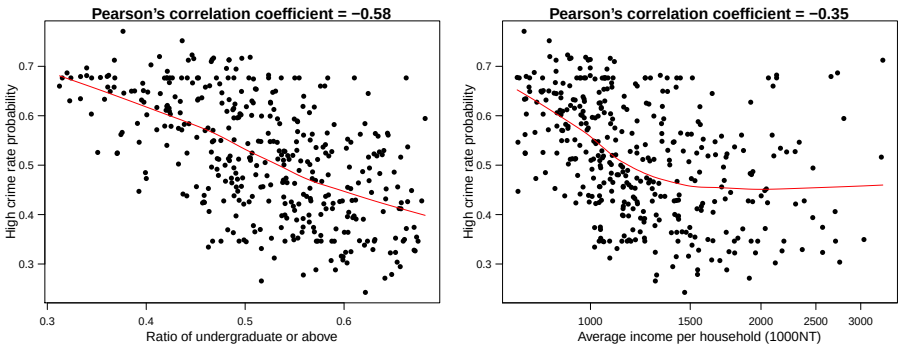

Fig. 6. Relationship between demographic and socioeconomic features and high crime rate probability.

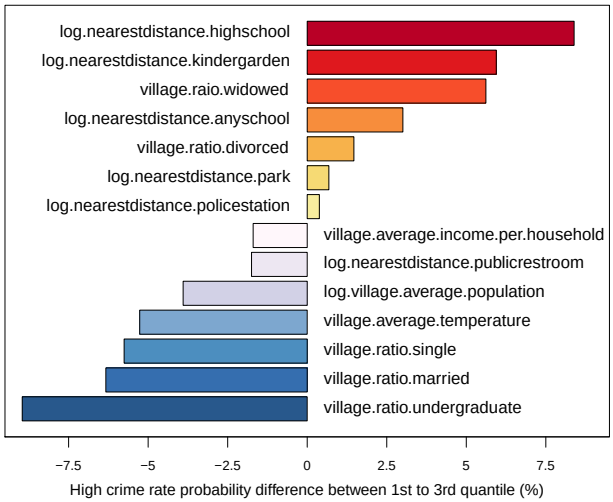

Fig. 7. The quantification of feature significance.

6 Conclusion and Future Work

This paper documents our efforts in applying data science techniques to a specific social problem of analyzing factors of high crime rates. The results illustrates the unique effectiveness of combing spatial data and demographic, social and economic factors. Going forward, we would like to investigate further into the mechanism of the SVM machine learning internals to gain better insights as to the inter-relationships between the selected key factors. Applying the same methodology to other cities/metro areas in Taiwan, or Greater China areas will help solidifying the modeling process and correctness. In addition, collaboration with other parts of the world will help us better understand if culture difference plays a role in crime factor analysis.

References

1. Open data from directorate-general of budget, accounting and statistics. http://www.dgbas.gov.tw/mp.asp?mp=1.
2. Open data ministry of interior. http://data.moi.gov.tw/MoiOD/default/Index.aspx.
3. Open data taipei. http://data.taipei/.
4. Supporting toolset for gis applications. http://gis.rchss.sinica.edu.tw/ISTIS/tools/.
5. Taiwan government open data. http://data.gov.tw/.
6. M. Anwer, S. Nasreen, and A. Shahzadi. Social and demographic determinants of crime in pakistan: A panel data analysis of province punjab. *International Journal of Economics*, 2015.
7. A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pages 427–434. ACM, 2014.
8. K. Bowers, M. Newton, and R. Nutter. A gis-linked database for monitoring repeat domestic burglary. *Mapping and Analysing Crime Data–Lessons from Research and Practice*, pages 120–137, 2001.
9. G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.
10. S. Chainey, L. Tompson, and S. Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1-2):4–28, 2008.
11. Y.-W. Chang and C.-J. Lin. Feature ranking using linear svm. In *WCCI Causation and Prediction Challenge*, pages 53–64, 2008.
12. H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau. Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56, 2004.
13. H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platts probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.
14. L. Lochner and E. Moretti. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *The American Economic Review*, 94(1):155–189, 2004.
15. A. T. Murray, I. McGuffog, J. S. Western, and P. Mullins. Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment. *British Journal of criminology*, 41(2):309–329, 2001.
16. J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
17. J. H. Ratcliffe and M. J. McCullagh. Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems*, 1(4):385–398, 1999.
18. V. Spicer, J. Song, P. Brantingham, A. Park, and M. A. Andresen. Street profile analysis: a new method for mapping crime on major roadways. *Applied Geography*, 69:65–74, 2016.
19. G. Sweeten, A. R. Piquero, and L. Steinberg. Age and the explanation of crime, revisited. *Journal of Youth and Adolescence*, 42(6):921–938, 2013.
20. N. Tilley, A. Tseloni, and G. Farrell. Income disparities of burglary risk security availability during the crime drop. *British Journal of Criminology*, page azr010, 2011.
21. T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Learning to detect patterns of crime. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer, 2013.