On Predicting Crime with Heterogeneous Spatial Patterns: Methods and Evaluation

Chuanxiu Xiong¹, Ajitesh Srivastava¹, Rajgopal Kannan², Omkar Damle³, Viktor Prasanna¹, Erroll Southers¹

¹ University of Southern California, Los Angeles, CA

Email: {chuanxix, ajiteshs, prasanna}@usc.edu, southers@price.usc.edu

² US Army Research Lab-West, Los Angeles, CA

Email: rajgopal.kannan.civ@mail.mil

³ New York University, New York City, New York

Email: ovd208@nyu.edu

ABSTRACT

Accurate prediction of crime incidents can assist the police in better planning of prevention strategies and scheduling deployment. The problem is often studied as a spatio-temporal regression problem approached by dividing the area of interest into a grid of uniform cells, and performing regression on timeseries of each cell. We propose that changing the method of division of the area can significantly improve crime prediction. We demonstrate this using a heterogeneous division of the area obtained by our partitioning algorithm that takes into account the density of crime. We further show that existing measures do not provide a fair comparison of two methods that partition the area in two different ways. To address this severe drawback in crime prediction evaluation, we propose a novel measure which is based on optimal allocation of resources relying on the prediction and then checking the actual number of crimes that would have been avoided by the allocation. Essentially, our measure answers the question of which model would have assisted in preventing most number of actual crimes if allocation were to be done using the predicted crimes. We also prove that a greedy algorithm results in the optimal allocation resources, thus making our evaluation computationally lightweight. Experiments on real-world datasets demonstrate that heterogeneous division of the area results in improved crime prediction while drastically decreasing the number of models to be trained compared to uniform grid division.

CCS CONCEPTS

• **Information systems** → **Temporal data**; *Spatial-temporal systems*; • **Computing methodologies** → **Supervised learning by regression**.

KEYWORDS

crime prediction; regression; spatial-temporal evaluation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6909-1/19/11...\$15.00

<https://doi.org/10.1145/3347146.3359374>

1 INTRODUCTION

In the past few years, the concept of ‘Predictive Policing’ has gained attention and agencies have started using crime prediction software¹. Many studies focus on advancing the learning approach in the crime predictive models based on additional features, such as text mining [1], demographic data [22], spatial-temporal relations [19], social media data [12, 30], or multiple data sources combined [2, 25, 29]. However, few efforts are made to address the effect of spatial partitioning of data.

In uniform grid division, the area size of each cell becomes crucial to the predictive model. When the area of a region is too small, the temporal data obtained from it may be too noisy for learning any pattern [16]. An extreme example would be when the region has been reduced to a cell that only contains 0 or 1 crime each week, the prediction model might not be able to predict the crimes in that cell accurately. Further, this also leads to training of many timeseries (one for each small cell), which can be computationally expensive. On the other hand, if the cells are made large, the specificity of the crime model suffers. An accurate prediction on a large area is of little importance to the police as no particular sub-area is specified. Therefore, there are two conflicting objectives: we wish to capture the trend in the data reliably while being specific in terms of the area. However, it is hard to choose a uniform grid size for the entire area of interest because crimes are not uniformly distributed over the map [4, 24]. For instance, in Figure 1, notice that some areas are dense while some are sparse depending on the opportunity of crime and physical environment.

Therefore, we propose a heuristic approach that divides the area into heterogeneous regions based on the number of crimes. The motivation is twofold: (i) improving crime prediction by capturing the heterogeneous patterns, (ii) and reducing training cost as multiple cells are merged into one heterogeneous region. Our hypothesis is that a sensible division of the area into heterogeneous regions can lead to “better” crime predictions. Nevertheless, judging which prediction is “better” is a difficult problem itself. Most studies use traditional evaluation metrics such as Root Mean Square Error (RMSE) [19, 25], Area Under Curve (AUC) [12, 31], Weighted Average Percentage Error (WAPE), etc. However, the traditional evaluation metrics are not designed to compare two methods that divide the area of interest differently, which change the timeseries themselves. A detailed description of the drawbacks of the existing

¹For example <http://www.predpol.com/about/>

Figure 1: Visualization of crime around University of Southern California: Distribution of crime is non-uniform. Different colors represent clusters.

evaluation metrics is presented in Section 2.2. Therefore, we also propose a metric which is based on measuring the number of crimes that would be avoided in the ground truth, if police resources were optimally distributed based on the predictions.

Specifically, our contributions are as follows:

- We demonstrate on real-world datasets that division of the area of interest into heterogeneous regions leads to “better” crime predictions with less computation cost as compared to traditional uniform division into grids.
- We propose a resource allocation based metric for evaluation of prediction models that overcomes the drawbacks of traditional spatio-temporal crime evaluation methods.

2 RELATED WORK

2.1 Related Work in Crime Prediction

Several existing work on crime prediction are based on uniform grid division [19, 28, 32]. Besides, existing studies suggest the use of regression and Neural Network for crime prediction, including Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) [7, 8, 11, 28]. In our work, we show that heterogeneous division of the area of interest produces better prediction results.

Besides, the studies in crime hotspots suggest the necessity of replacing the traditional uniform grid division in crime prediction. Crime hotspots are commonly referred to as areas where the concentration of crime is higher [9] and have been thoroughly studied in the past decades [26, 27]. The recent studies generally focus on better identifying hotspots [17] and focusing police resources [10, 20]. However, the studies do not account for the regression problem which is our focus, i.e., predicting the number of incidents in a small region, instead of just predicting which regions are hot-spots.

2.2 Drawbacks of Existing Metrics

Even though there are existing evaluation metrics that compare different crime prediction models, these metrics have inherent flaws when they are applied to crime prediction models that divide the area of interest differently. Here, we list three popular metrics: Area Under Curve (AUC) [3], Root Mean Squared Error (RMSE) and Weighted Average Percentage Error (WAPE) [6] and show

why these metrics are not suitable for different divisions of area of interest.

Area Under Curve (AUC) represents the area under the receiver operating characteristic curve, i.e. ROC curve [3]. AUC in the context of crime prediction is the area under the surveillance curve - % incidents covered vs % area of surveillance. Suppose the area is divided into three regions of equal area with ground truth incidents (30, 2, 0). Assume two models predict (25, 2, 0) and (10, 1, 0). Both will be assigned the same AUC score while, clearly, the first prediction is “better”.

RMSE and WAPE are defined as the following

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (A_i - F_i)^2}{n}} \quad (1)$$

$$WAPE = \frac{\sum_{i=1}^n |A_i - F_i|}{\sum_{i=1}^n A_i} \quad (2)$$

subject to $1 \leq i \leq n$ and $i \in \mathbb{Z}$,

where A_i is the actual number of crimes in region i , F_i is the forecasted number of crimes in region i , and n is the total number of regions. Let the region i be split into two regions: i_1 and i_2 and error for region i be e_i . Therefore, we have the following equations:

$$e_i = A_i - F_i \quad (3)$$

$$A_i = A_{i1} + A_{i2} \quad (4)$$

$$F_i = F_{i1} + F_{i2} \quad (5)$$

$$e_i = e_{i1} + e_{i2} \quad (6)$$

Notice that

$$e_i^2 = (e_{i1} + e_{i2})^2 \geq e_{i1}^2 + e_{i2}^2, \quad (7)$$

which means RMSE would decrease by splitting the heterogeneous regions. On the other hand, for WAPE, by triangle inequality, we have

$$|e_i| = |e_{i1} + e_{i2}| \leq |e_{i1}| + |e_{i2}|, \quad (8)$$

which means WAPE would decrease by merging the heterogeneous regions. RMSE is biased towards subdividing larger regions while WAPE is biased towards merging. Therefore, it is not fair to use either RMSE or WAPE to compare two prediction models that are based on different subdivisions of the area of interest.

3 METHODOLOGY

Our objective is to devise an algorithm for partitioning the area of interest into heterogeneous regions, using which we can demonstrate that doing so can improve crime prediction as compared to traditional uniform grids, when used in conjunction with regression and deep learning models. To evaluate the prediction, we propose a resource allocation metric which measures how many crimes can be stopped if police resources were optimally distributed to based on the predictions.

Figure 2 shows our methodology consisting of the following three steps: defining heterogeneous crime-prone regions, predicting crimes within each region, and evaluating the predictions. The traditional methods of crime prediction divide the area of interest into uniform grids, e.g., 2×2 , 8×8 , 16×16 , etc., as shown in Figure 3. To define heterogeneous regions, we start with fine grained uniform grids, such as 40×40 . Then, we perform a grid clustering to aggregate the grid cells into heterogeneous regions, the details

Figure 2: Overall architecture of the Methodology

Figure 3: Different Grid Shapes of One Area of Interest.

of which are introduced in the following subsection 3.1. With the new regions, we convert the crime data into weekly timeseries data and predict crimes individually per region. Finally, we evaluate the prediction using the metric we propose in the following subsection 3.2, which allows comparisons among different models with different grid shapes.

3.1 Grid Clustering Algorithm

We propose a simple grid clustering algorithm (Algorithm 1) to demonstrate that heterogeneous regions provide better predictions compared to uniform grids. Figure 4 illustrates our clustering approach, which shows the result when we run the algorithm on an arbitrary 3×3 grid with a threshold of 20. First, the algorithm sorts the cells into decreasing average number of crimes. All cells that are above or equal to the threshold 20 form their own region, such as B, F, and D. Next, A searches for the nearest densest neighbor that has not been included, which is E, and merges with and becomes equal to the threshold. Similarly, H merges with I and G. C is the only region that is left behind with number of crimes below threshold. If the number of the regions is n , an upper bound on the time complexity of the algorithm is $O(n \log n)$. Our goal is to show that heterogeneous regions lead to better prediction of crime compared

to uniform grids. The question of finding the “best” heterogeneous partitioning is not the focus in this paper, and is left for future work.

Algorithm 1 Naive grid clustering algorithm

```

1: procedure CLUSTERING(crimes, grid, threshold)
2:   for grid[i][j] in decreasing density ▶ Traverse the cells from
   the densest to the sparsest crime incidents
3:   if grid[i][j] has been visited then
4:     continue
5:   if crimes[i][j]  $\leq$  threshold then
6:     neighbors = {Moore-neighbors of grid[i][j] that have not
   been visited}
7:     while crimes[i][j]  $\leq$  threshold do
8:       if neighbors.length == 0 then
9:         break
10:      Merge the neighbor with most crimes to grid[i][j]
11:      Remove the neighbor from neighbors

```

3.2 Evaluation Metric

Once we have the predictions for all regions, we test the effectiveness of the predictions based on how many crime incidents can be

Figure 4: Region layout defined by Algorithm1 on a 3×3 grid with a threshold of 20.

avoided with optimal distribution of finite resources k given the predictions. The goal of the metric is only evaluating the prediction, not performing actual assignments as actual assignment may affect the crime patterns. Note that we are only given the predictions for the regions, and no finer grained details are known. Therefore, we assume that the incidents in a given region are uniformly distributed over it. Otherwise, the crime prediction method should have identified a further subdivision. Our resource allocation metric considers the area and the number of crimes in a region. It calculates the fraction of crimes captured with varying number of resources. This is done under certain assumptions of the effect of a police resource on crime.

For instance, suppose one police resource can successfully prevent n_0 number of crimes in area of m_0 square units. A higher number of crimes require more resources to address all crimes, and a larger area also requires more resources to cover the whole area. Effectiveness of a resource decreases as the area increases (inversely proportional). Besides, we assume one police resource cannot cross the boundaries. This favors more specific prediction of crime in terms of area. Due to the lack of knowledge of distribution within a selected area, we assume that the effect of a resource spreads uniformly over it. For instance, an area of 1 unit with $2n_0$ crimes will require 2 units. Similarly, an area of 2 units with n_0 crimes will require 2 units. For a region i with k_i police resources and a_i area in square units, the maximum number of crimes which can be prevented is $\frac{k_i m_0 n_0}{a_i}$. k_i is assigned based on the predicted number of crimes in region i , c_i . The problem becomes an optimization problem of maximizing

$$\sum_{i=1}^n k_i b_i \quad (9)$$

$$\text{subject to: } k_i b_i \leq c_i \text{ and } \sum_{i=1}^n k_i \leq k \quad (10)$$

where $b_i = \frac{m_0 n_0}{a_i}$, and k represents the total number of available units of police resources.

After a model has calculated its prediction values, the following steps need to be followed for calculating its effectiveness under the above assumptions:

- 1) Select the cluster region with minimum area a_i i.e. maximum b_i .
- 2) Assign as many resources possible to it subject to requirements.
- 3) Repeat the above two steps until the resources are finished or all the regions are covered.

This resource allocation approach is a greedy algorithm that can be proven to be optimal, i.e. an algorithm which prevents the most number of crimes provided all the predictions are accurate. More generally, a greedy algorithm applies as long as the effect of resources in the regions (utility) is dictated by concave functions. For the general setting, we propose the following problem that leads to the evaluation of the prediction models through resource assignment.

PROBLEM DEFINITION 1 (UNITS ASSIGNMENT PROBLEM). Given $k \in \mathbb{Z}$ resources and n concave non-decreasing utility functions $f_i : \mathbb{Z} \rightarrow \mathbb{R}$, where $f_i(z_i)$ represents the utility of assigning z_i units to function f_i , maximize the total utility $F = \sum_i f_i(z_i)$ subject to $\sum_i z_i = k$.

Algorithm 2 Greedy Maximization using Marginal Returns

```

1: function GREEDYMAX( $(f_1, f_2, \dots, f_n), k$ )
2:   for  $i \leftarrow 1 : n$  do
3:      $z_i \leftarrow 0$ 
4:   for  $j \leftarrow 1 : k$  do
5:      $idx \leftarrow \arg \max_i (f(z_i + 1) - f(z_i))$ 
6:      $z_{idx} \leftarrow z_{idx} + 1$ 
7:   return  $(z_1, z_2, \dots, z_n)$ 

```

LEMMA 3.1. For a non-decreasing concave function $f : \mathbb{Z} \rightarrow \mathbb{R}$, and $h \geq 1$,

$$f(x + h) - f(x) \leq h(f(x) - f(x - 1)) \quad (11)$$

$$f(x) - f(x - h) \geq h(f(x) - f(x - 1)) \quad (12)$$

We prove the following.

THEOREM 3.2. Algorithm 2 produces the optimal assignment for Units Assignment Problem.

PROOF. Suppose the greedy assignment results in an assignment of z_i to the function f_i . Without loss of generality, we assume that the functions are ordered as following: if $i < j$ then $f_i(z_i) - f_i(z_i - 1) \geq f_j(z_j) - f_j(z_j - 1), \forall i, j$ such that $z_i, z_j \geq 1$.

Assume that the optimal assignment $(z_1^*, z_2^*, \dots, z_n^*)$ is different from greedy assignment and produces a greater $F = \sum_i f_i(z_i^*)$.

Choose the smallest index p such that $f_p(z_p) \neq f_p(z_p^*)$.

Case $f_p(z_p) > f_p(z_p^)$.* Since, $\sum_i z_i = \sum_i z_i^* = k, \exists p < i_1 < i_2 < \dots < i_M$, for some $M > 0$ such that $z_{i_r}^* > z_{i_r}$ and $\sum_r (z_{i_r}^* - z_{i_r}) \geq z_p - z_p^*$. Therefore, it is possible to pick $h_{i_r} \leq (z_{i_r}^* - z_{i_r})$ such that $\sum_r h_{i_r} = z_p - z_p^*$. Suppose we take h_{i_r} out of the optimal assignment $z_{i_r}^*, \forall r$ and assign them to the function f_p , then we

should not expect any gain ($\Delta F \leq 0$) as the assignment we started with was optimal. We note that

$$\begin{aligned}
& \sum_r (f_{i_r}(z_{i_r}^*) - f_{i_r}(z_{i_r}^* - h_{i_r})) \\
& \leq \sum_r h_{i_r} (f_{i_r}(z_{i_r}^* - h_{i_r}) - f_{i_r}(z_{i_r}^* - h_{i_r} - 1)) \\
& \text{[Using Lemma 3.1]} \\
& \leq \sum_r h_{i_r} (f_{i_r}(z_{i_r}) - f_{i_r}(z_{i_r} - 1)) \text{ [Due to concavity]} \\
& \leq \sum_r h_{i_r} (f_p(z_p) - f_p(z_p - 1)) \text{ [Due to ordering of functions]} \\
& \leq (z_p - z_p^*) (f_p(z_p) - f_p(z_p - 1)) \text{ [Since } \sum_r h_{i_r} = z_p - z_p^*] \\
& \leq (f_p(z_p) - f_p(z_p^*)) \text{ [Using Lemma 3.1]}
\end{aligned}$$

Therefore, the gain obtained in this case is $\Delta F = (f_p(z_p) - f_p(z_p^*)) + \sum_r (f_{i_r}(z_{i_r}^* - h_{i_r}) - f_{i_r}(z_{i_r}^*)) \geq 0$.

Case $f_p(z_p) < f_p(z_p^)$.* Since, $\sum_i z_i = \sum_i z_i^* = k$, $\exists p < i_1 < i_2 < \dots < i_M$, for some $M > 0$ such that $z_{i_r} - z_{i_r}^*$ and $\sum_r (z_{i_r} - z_{i_r}^*) \geq z_p - z_p^*$. Therefore, it is possible to pick $h_{i_r} \leq (z_{i_r} - z_{i_r}^*)$ such that $\sum_r h_{i_r} = z_p^* - z_p$. Now, we take $z_p^* - z_p$ resources out of the optimal assignment on f_p and distribute them such that f_{i_r} gets $z_{i_r} + h_{i_r}$. Since we have obtained the sequence $z_1, z_2, \dots, z_n$ using greedy assignment, $f_i(z_i + 1) - f_i(z_i) \leq f_j(z_j) - f_j(z_j - 1)$, $\forall i \neq j$, otherwise the last unit that went to f_j would have gone to f_i instead. We have $f_p(z_p^*) - f_p(z_p)$

$$\begin{aligned}
& \leq (z_p^* - z_p) (f_p(z_p + 1) - f_p(z_p)) \text{ [Using Lemma 3.1]} \\
& \leq \sum_{i_r} h_{i_r} (f_p(z_p + 1) - f_p(z_p)) \text{ [Since } \sum_r h_{i_r} = z_p - z_p^*] \\
& \leq \sum_{i_r} h_{i_r} (f_{i_r}(z_{i_r}) - f_{i_r}(z_{i_r} - 1)) \text{ [From of Algorithm 2]} \\
& \leq \sum_{i_r} h_{i_r} (f_{i_r}(z_{i_r}^* + h_{i_r}) - f_{i_r}(z_{i_r}^* + h_{i_r} - 1)) \\
& \leq \sum_{i_r} (f_{i_r}(z_{i_r}^* + h_{i_r}) - f_{i_r}(z_{i_r}^*)) \text{ [Using Lemma 3.1]} \\
& \leq \sum_{i_r} (f_{i_r}(z_{i_r}) - f_{i_r}(z_{i_r}^*)) \text{ [} f_{i_r} \text{ is non-decreasing].}
\end{aligned}$$

Therefore, $\Delta F = \sum_{i_r} h_{i_r} (f_{i_r}(z_{i_r}) - f_p(z_p^*)) + (f_p(z_p) - f_p(z_p^*)) \geq 0$.

But $\Delta F \leq 0$, and so ΔF must be zero, i.e., for any optimal assignment that differs from the greedy assignment first at index p , we can perform a reassignment that retains optimality so that they no longer differs at index p . Proceeding thus we get $z_i^* = z_i$, $\forall i$. Hence, the greedy assignment is optimal. $\square$

4 EXPERIMENTS

4.1 Datasets

We conducted our experiments on two datasets:

DPS. This is the University of Southern California main campus crime data from February 10, 2010 to August 4, 2016 obtained from Department of Public Safety at USC. The range of latitude and longitude are: 34.015° N to 34.037° N and 118.302° W to 118.27° W. There are 24674 crime incidents.

LA. This is the Los Angeles crime data² from January 1, 2014 to December 31, 2016. The range of latitude and longitude are: 33.781° N to 34.129° N and 118.435° W to 117.581° W. There are 381934 crime incidents. This area was specifically chosen to form a rectangular area of interest that does not contain ocean.

We use the latitude, longitude, and time of crime incidents from each of the datasets.

4.2 Experiments

We evaluate different uniform grids as well as heterogeneous divisions with different thresholds. The uniform grid configurations are represented as $n \times m$, such as 2×2 , 8×8 . Besides choosing the parameters, each experiment consists of three steps: defining heterogeneous regions, predicting the crimes and evaluating the predictions. We define the first two-thirds of the data as the training data and the rest as the testing data. We use the training data to define heterogeneous regions and train prediction models within each region. We use the testing data to evaluate the prediction results. Using our evaluation metric, we take the traditional uniform region prediction as the baseline approach and compare it with our heterogeneous region prediction based on Algorithm 1.

4.2.1 Defining Heterogeneous Regions. Based on the grid shapes and thresholds, we define the heterogeneous regions. We first retrieve the maximum and minimum values of both latitude and longitude, combination of which defines a rectangle where all the crimes have happened. This is the area of interest. We uniformly divide the region into small uniform cells and assign each incident to the cell it belongs to based on its location. We cluster using Algorithm 1 to obtain regions, convert the data into weekly data for each region and proceed to crime predictions.

4.2.2 Crime Predictions. We explore the effect of heterogeneous regions on three prediction methods:

Moving Mean (MM). MM model with a window q predicts the next data point as the mean of the previous q values. For MM method, we perform a grid search on q with cross-validation and mean squared error as the measure of fitting. Even though mean squared error cannot be used as a metric for the overall prediction, we utilize it to select the appropriate hyperparameter q because it is the traditional method.

Autoregressive Integrated Moving Average (ARIMA). ARIMA model includes an Autoregressive model with a window p , which applies a regression on the previous p data points to predict the next data point and a Moving Average Model (MA) that predicts the next data point as the mean of the current and previous q white noise terms [14, 18]. In our experiment, we consider the seasonality because crime (by type) is a seasonal phenomenon (violent crime increases in summer, thefts and robberies during Christmas, etc.) [13]. We assume the seasonality of the data to be 52 weeks. For ARIMA method, we choose the optimal hyperparameters, train an independent prediction model and predict crimes with different forecasting horizons for each cell/heterogeneous region. To choose appropriate hyperparameters, we first conduct Augmented Dickey-Fuller [21] and Canova-Hansen [5] to estimate the optimal order of differencing, d and the order of seasonal differencing, D . Given the

²<https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq>

orders of differencing and seasonal differencing, we fit the model and utilize Akaike information criterion (AIC) and Bayesian information criterion (BIC) [23] as estimators of the relative quality of the model. We use gradient-based optimization to optimize the hyperparameters.

Long Short-Term Memory (LSTM). LSTM network is a Recurrent Neural Network with LSTM units. In our architecture, we use LSTM with a forget gate, which consists of a cell, an input gate, an output gate and a forget gate. [15]

After finding an optimal prediction model in a region, we conduct an n -step-ahead prediction with n being different forecasting horizons. We choose n to be 1, 2, 4, 8, meaning forecasting crimes 1 week, 2 weeks, 4 weeks, and 8 weeks ahead, respectively.

We need to find an individual set of hyperparameters for each region because each region could be very different from one another due to the uneven crime distribution. As a result, finding the optimal hyperparameters could be computationally expensive and largely dependent on how many regions we have. Besides, the process of defining heterogeneous regions and finding hyperparameters have to be repeated regularly to accommodate the changing crime patterns responsive to the police resources. Based on the two reasons above, it is crucial to have fewer regions (and hence, fewer models) while maintaining the quality of the predictions.

4.2.3 Evaluations. We evaluate the predictive models by increasingly assigning police resources to the area and applying the evaluation metric. We keep assigning resources until we avoid all the crimes in every combination of grid shapes and thresholds. In this paper, we present a selected set of plots as the remaining results in similar trends.

4.3 Results and Discussion

In this section, we present the evaluation results with forecasting horizons of 1 week and 8 weeks. In the plots (Figures 5, 6, 7, and 8), the y-axis represents the fraction of crimes that are avoided based on our evaluation metrics; the x-axis represents the total units of resources we apply to the area of interest. Each line represents one experiment with either uniform grid (i.e. 2×2 , 8×8 , etc.) or heterogeneous division with threshold and an initial grid shape (i.e. 1750, 2250, etc.). As described in the previous section, every line has an initial grid shape as well as a clustering threshold. The lines with a threshold 0 means predictions with uniform regions, which are our baseline results. In Tables 1, 2, 3 and 4, we list the number of models we have to train in different configurations.

The experiments support the credibility of our naive clustering algorithm and evaluation metric. To help understand the results, there are several points that we need to emphasize:

4.3.1 Baseline methods and our methods. Since we initialize our heterogeneous division with a uniform division, a threshold of 0 would output uniform grids. Therefore we denote the baseline methods as $m \times n(0)$. We are comparing the computation cost during training (the number of models we need to train) as well as the performance of each configuration.

4.3.2 The computation cost of prediction. Since each heterogeneous region is a separate model that we need to train, the cost of prediction is high, especially with more complicated methods such

as ARIMA and LSTM. Meanwhile, model retraining needs to be performed repetitively in reality as the crime patterns will be influenced by the police resources. Therefore, it is important to consider how clustering could help reduce the computation cost by reducing the number of models that we need to separately train.

4.3.3 The scale of evaluation metrics. Since we abstracted our resources as k_i , the effect of a police resource, and the area a_i , the actual difference between the evaluation curves may get scaled by the choices of these parameters.

4.3.4 Experimental nuance between DPS and LA data set. To better present the results, we chose different hyperparameters in the evaluation metric for each data set. Besides, many details that could affect the interpretation of the results are different in both data sets, such as the length of the time series, the area of interest, the distribution of the crimes, etc. Therefore, the results of each data set are intended to be viewed independently.

Table 1: Number of clusters/models on DPS data under different thresholds based on 225-week training data (Grid shape: 20×20)

Threshold	Clusters
0	400
200	61
400	39
800	22
1200	14
2000	9

Table 2: Number of clusters/models on LA data under different thresholds based on 104-week training data (Grid shape: 40×40)

Threshold	Clusters
0	1600
2000	99
4000	58
6000	40
8000	31
10000	24

Table 3: Comparison of number of models against the best performing baseline on DPS dataset

Grid shape	Threshold	Clusters
8×8	0	64
20×20	1750	11
20×20	2000	9
20×20	2250	8

The first observation is that we are able to reduce the computation cost while maintaining a high accuracy of crime prediction

Figure 5: 1-week-ahead Crime Prediction on the DPS data.

Figure 6: 8-week-ahead crime prediction on the DPS data.

Figure 7: 1-week-ahead crime prediction on the LA data.

by applying clustering. This is very important since the crime prediction needs to be repeated regularly as the presence of police

resources will shift crime patterns. Table 1 and 2 show how clustering could drastically decrease the number of heterogeneous regions

Figure 8: 8-week-ahead crime prediction on the LA data.

Table 4: Comparison of number of models against the best performing baseline on LA dataset

Grid shape	Threshold	Clusters
16×16	0	256
40×40	9500	26
40×40	10000	24
40×40	10500	24

- the number of models that we need to train. In Table 1, the number starts as 400 when no clustering is applied and drops to 9 when the clustering threshold is around 2000. The computation cost is further lowered in the LA dataset as the number of clusters drops from 1600 to 24 at the threshold of 10000. In both scenarios, the number of models that we need to train is reduced drastically by more than 99%. Even when compared to the best uniform grid predictions, the heterogeneous patterns give us considerable improvement on run-time. Table 3 and 4 list the number of clusters of the best uniform grid predictions (8×8 in DPS dataset and 16×16 in LA dataset) as well as the competing heterogeneous predictions. For example, the number of clusters drops by 87.5% (from 64 to 8 at the threshold of 2250) in DPS dataset and by 90.625% (from 256 to 24) in LA dataset. These results show a potential of providing a better, fine-grained based prediction result without taking the computation cost.

In terms of the number of crimes avoided, the heterogeneous spatial patterns with appropriate threshold values beat the baseline approaches. Figure 5 and 6 show that on the DPS dataset, the clustering with threshold around 2000 and gridshape 20×20 performs exceptionally well compared to other baseline results, beating the best baseline result (8×8 with threshold 0) in all three methods. The baseline and the clustered results together bolster our perception of the problem that is stated in Section 1 – the gridshapes with few cells perform poorly (i.e. gridshape 2×2) because the specificity of the model suffers; but if the area is too finely divided (i.e. gridshape 16×16 in DPS), the prediction result would suffer because there is too little information inside each cell. Such conflict gives the baseline method a bottleneck. The clustered results listed here are derived from a finer grid division (i.e. gridshape 20) with

appropriate clustering threshold. We also observe that the performance of heterogeneous division does not get affected significantly by the choice of threshold parameter.

Similar patterns could be seen in the LA dataset (Figure 7 and 8), where the clustering with threshold 10000 and gridshape 40×40 are better than other baseline results. We also see very little difference between 1-week-ahead and 8-week-ahead predictions. This shows that crime patterns in heterogeneous regions are relatively stable and thus more predictable over time.

The only exception occurs in ARIMA prediction, where the clustered results exceed the baseline most of the time except when there are abundant resources. In fact, by comparing the different prediction methods (MM, ARIMA, LSTM) under different configurations, we found traditional regression models such as MM and ARIMA are more sensitive to the dataset and parameters, while LSTM constantly gives the best results among the three methods. Some of the comparisons are shown in Figure 9. It can be seen that the performances of ARIMA and MM drops as the grid configuration becomes less ideal for crime prediction. For example, in Figure 5, the MM result is very close to LSTM when the threshold and configuration are appropriate (gridshape: 20×20 , threshold: 1750) but starts to drop as the gridshape and threshold changes. Similar situations happen in Figure 7, where the ARIMA gives results that are far worse than LSTM when the gridshape and threshold change to a less desirable state. In all results, LSTM performs constantly better or as good as MM and ARIMA. This provides insight on how the deep learning method performs compared to regression based methods on spatial-temporal prediction.

5 CONCLUSION

We have addressed the problem of improving data-driven prediction of crime. Particularly, we have shown that heterogeneous division of area of interest results in better prediction of crimes, while requiring fewer models to be trained. We have also addressed the drawbacks in traditional spatio-temporal prediction evaluation metrics in this domain, and provided a novel evaluation of crime prediction based on using the prediction to prevent crimes. It assumes the predictions to be perfect and assigns resources optimally

Figure 9: 1-week-ahead crime prediction with different algorithms

based on them. Then it computes how many crimes would have been stopped by considering the ground truth data. Experiments on real-world datasets demonstrated significant improvement through heterogeneous division, in terms of number of models required to be trained and sensitivity to parameters. In future work, we will consider the effect of other clustering algorithms for improving crime prediction.

ACKNOWLEDGMENTS

This work is supported by U.S. National Science Foundation under EAGER Award No.:1637372.

REFERENCES

- [1] Meshrif Alruily. 2012. Using text mining to identify crime patterns from arabic crime news report corpus. (2012).
- [2] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*. ACM, 427–434.
- [3] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.
- [4] Marilyn A Brown. 1982. Modelling the spatial distribution of suburban crime. *Economic Geography* 58, 3 (1982), 247–261.
- [5] Fabio Canova and Bruce E Hansen. 1995. Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business & Economic Statistics* 13, 3 (1995), 237–252.
- [6] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions* 7 (2014), 1525–1534.
- [7] Peng Chen, Hongyong Yuan, and Xueming Shu. 2008. Forecasting crime using the arima model. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 5. IEEE, 627–630.
- [8] Bitzel Cortez, Berny Carrera, Young-Jin Kim, and Jae-Yoon Jung. 2018. An architecture for emergency event prediction using LSTM recurrent neural networks. *Expert Systems with Applications* 97 (2018), 315–324.
- [9] John Eck, Spencer Chainey, James Cameron, and Ronald Wilson. 2005. Mapping crime: Understanding hotspots. (2005).
- [10] Emre Eftelioglu, Shashi Shekhar, and Xun Tang. 2016. Crime hotspot detection: A computational perspective. In *Data Mining Trends and Applications in Criminal Science and Investigations*. IGI Global, 82–111.
- [11] HICHAM AIT EL BOUR, Soumaya Ounacer, Yassine Elghomari, Houda Jihal, and Mohamed Azzouazi. 2018. A crime prediction model based on spatial and temporal data. *Periodicals of Engineering and Natural Sciences* 6, 2 (2018), 360–364.
- [12] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.
- [13] Wilpenm Gorr, Andreas Olligschlaeger, and Yvonne Thompson. 2003. Short-term forecasting of crime. *International Journal of Forecasting* 19, 4 (2003), 579–594.
- [14] James Douglas Hamilton. 1994. *Time series analysis*. Vol. 2. Princeton university press Princeton, NJ.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. DOI: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>
- [16] Hyeon-Woo Kang and Hang-Bong Kang. 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PLoS one* 12, 4 (2017), e0176244.
- [17] Nick Malleson and Martin A Andresen. 2015. Spatio-temporal crime hotspots and the ambient population. *Crime science* 4, 1 (2015), 10.
- [18] Richard McCleary, Richard A Hay, Erroll E Meidinger, and David McDowall. 1980. *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA.
- [19] Michail Misyrilis, Chung Ming Cheung, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. 2017. Spatio-Temporal Modeling of Criminal Activity. In *Proceedings of the 2nd International Workshop on Social Sensing*. ACM, 3–8.
- [20] George Mohler. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting* 30, 3 (2014), 491–497.
- [21] Rizwan Mushtaq. 2011. Augmented dickey fuller test. (2011).
- [22] E Britt Patterson. 1991. Poverty, income inequality, and community crime rates. *Criminology* 29, 4 (1991), 755–776.
- [23] David Posada and Thomas R Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology* 53, 5 (2004), 793–808.
- [24] Jerry H Ratcliffe and Michael J McCullagh. 1999. Hotbeds of crime and the search for spatial accuracy. *Journal of geographical systems* 1, 4 (1999), 385–398.
- [25] Ginger Saltos and Mihaela Cocca. 2017. An exploration of crime prediction using data mining on open data. *International Journal of Information Technology & Decision Making* 16, 05 (2017), 1155–1181.
- [26] Lawrence W Sherman, Patrick R Gartin, and Michael E Buerger. 1989. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology* 27, 1 (1989), 27–56.
- [27] William Spelman. 1995. Criminal careers of public places. *Crime and place* 4 (1995), 115–144.
- [28] Alexander Stec and Diego Klabjan. 2018. Forecasting Crime with Deep Learning. *arXiv preprint arXiv:1806.01486* (2018).
- [29] Xiaofeng Wang, Donald E Brown, and Matthew S Gerber. 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *2012 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 36–41.
- [30] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 231–238.
- [31] Senuri Wijenayake, Timothy Graham, and Peter Christen. 2018. A Decision Tree Approach to Predicting Recidivism in Domestic Violence. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 3–15.
- [32] Xiangyu Zhao and Jiliang Tang. 2017. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 497–506.