Pervasive and Mobile Computing 53 (2019) 62–74

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments

Charlie Catlett [a,b], Eugenio Cesario [c,d,*], Domenico Talia [e], Andrea Vinci [d]

[a] University of Chicago, USA
[b] Argonne National Laboratory, USA
[c] Monmouth University, USA
[d] ICAR-CNR, Italy
[e] University of Calabria, Italy

ARTICLE INFO

Article history:
Available online 17 January 2019

Keywords:
Crime prediction
Smart city
Urban computing
Data analytics

ABSTRACT

Steadily increasing urbanization is causing significant economic and social transformations in urban areas, posing several challenges related to city management and services. In particular, in cities with higher crime rates, effectively providing for public safety is an increasingly complex undertaking. To handle this complexity, new technologies are enabling police departments to access growing volumes of crime-related data that can be analyzed to understand patterns and trends. These technologies have potentially to increase the efficient deployment of police resources within a given territory and ultimately support more effective crime prevention. This paper presents a predictive approach based on spatial analysis and auto-regressive models to automatically detect high-risk crime regions in urban areas and to reliably forecast crime trends in each region. The algorithm result is a spatio-temporal crime forecasting model, composed of a set of crime-dense regions with associated crime predictors, each one representing a predictive model for estimating the number of crimes likely to occur in its associated region. The experimental evaluation was performed on two real-world datasets collected in the cities of Chicago and New York City. This evaluation shows that the proposed approach achieves good accuracy in spatial and temporal crime forecasting over rolling time horizons.

1. Introduction

Reference Context. The 21st Century is frequently referenced as the "Century of the City", reflecting the unprecedented global migration into urban areas that is under way [1,2]. This steadily increasing urbanization is bringing vexing social, economic, and environmental transformations to urban areas. For example, it is presenting challenges to organizations tasked with city management and provision of essential services, like resource planning (water, electricity), transit, air and water quality, and public safety [3]. Moreover, for cities with higher crime rates, crime spiking is becoming one of the most important social problems, affecting not only public safety but also health, education, child development, and adult socio-economic status [4,5].

Motivations and Contributions. An ever-increasing volume of urban-related data, with spatial and temporal attributes, from weather to air quality to economic activity, is available for public organizations, including police departments, to

* Corresponding author.
E-mail addresses: ecesario@monmouth.edu, cesario@icar.cnr.it (E. Cesario).

https://doi.org/10.1016/j.pmcj.2019.01.003
1574-1192/

integrate with internal data. This offers the opportunity to apply data analytics methodologies to extract useful predictive models related to crime events, which can enable police departments to better utilize their limited resources and develop more effective strategies for crime prevention. In particular, extensive criminal justice studies show that the incidence of criminal events is not equally distributed within a city. In fact, crime rates can change with respect to the geographic location of the area (there are low-risk and high-risk areas) and crime trends can vary (seasonal patterns, peaks, dips) with respect to the period of the year. For this reason, an accurate predictive model must be able to automatically detect both which areas in the city are more affected by crime events and how the crime rate of each specific area varies with respect to the temporal period. This knowledge can enable police departments to efficiently allocate their resources to specific crime hot spots, allowing for the effective deployment of officers to areas of high risk or removal of officers from areas seeing decreasing levels of crime, thus more efficiently preventing or quickly responding to criminal activity.

This paper presents the design and implementation of an approach based on spatial analysis and auto-regressive models to automatically detect high-risk crime regions in urban areas and to reliably forecast crime trends in each region. The algorithm is composed of several steps. First, high crime density areas (called crime dense regions, or crime hotspots) are discovered through a spatial analysis approach, where shapes of the detected regions are automatically traced by the algorithm without any pre-fixed division in areas. Then, a specific crime prediction model is discovered from each detected region, analyzing the partitions discovered during the previous step. The final result of the algorithm is a spatio-temporal crime forecasting model, composed of a set of crime dense regions and a set of associated crime predictors, each one representing a predictive model to forecast the number of crimes that are estimated to happen in its specific region.

As a case study, we present here the analysis of crimes within (i) a large area of Chicago and (ii) the borough of Manhattan in New York City, involving about two million crime events (over a period of 16 yr) and 1.5 million crime events (over a period of 11 yr), respectively. Chicago crime data has been gathered by the Plenario platform [6], a Web framework that provides public access to more than one hundred urban datasets, while the New York City crime data has been gathered from the New York City Opendata platform [7]. The results of the experimental evaluation show the effectiveness of the approach, by achieving good accuracy in spatial and temporal crime forecasting over rolling time horizons. We also present a comparative analysis of the results obtained through our approach with other algorithms presented in the literature, demonstrating higher accuracy of the proposed algorithm relative to other regressive approaches proposed in literature. For the sake of clarity, this paper extends the work presented in [8] and it provides several original contributions with respect to the previous one. The most significant extension concerns the experimental evaluation in Section 5, which has been extended by testing the proposed algorithm on a second real-world case study (New York City), and by performing a comparative analysis with other regression analysis approaches proposed in literature.

Plan of the Paper. The rest of the paper is organized as follows. Section 2 reports the most important approaches in crime data mining literature and the most representative projects in that field of research. Section 3 outlines the problem statement and goals of our analysis. Section 4 presents the Spatio-Temporal Crime Prediction algorithm by describing its steps in detail. Section 5 describes the experimental evaluation, performed on two real-world case studies. Finally, Section 6 concludes the paper and plans future research works.

2. Related work

Several data mining techniques have been used for crime analysis. Some approaches have been proposed for *crime location prediction* [4,9], while others are aimed at *crime pattern detection* [10–13]. In this section we briefly review the most representative research work in both the areas. Then, we report a critical comparison (on the basis of some specific features) among the method we developed and state-of-art solutions.

Crime location prediction. CrimeTracer [4] is based on a probabilistic framework to model the spatial behavior of known offenders within areas they frequent, called *activity spaces*. Experiments carried out on real-world crime data have shown that criminals frequently commit crimes within their activity spaces, rather than venture into unknown territories. The authors in [9] model crime location predictions as a special case of spatial data mining classification task, and exploit one-class support vector machines (SVM) to classify locations as hot-spot or no hot-spot crime areas.

Crime pattern detection. The approach proposed in [10] exploits Negative Binomial Regression to infer crime rates in different city areas, integrating geographic, demographic, POIs and taxi flows data. Multivariate time series clustering and ARIMA models are proposed in [12] and [14], to discover similar crime trends and to make short-term forecasting of crimes, respectively. Recurrent Neural Networks models, which exploits spatial and temporal information for forecasting crime hotspots, are presented in [11]. In [13] Holt Exponential Smoothing has been experimented using city-wide data and resulted as an accurate forecast model for precinct-level crime series.

Table 1 reports a more detailed and critical comparison among the proposed approach and some other solutions proposed in the literature. The comparison takes into account four features, as detailed in the following.

Crime hotspot detection. This feature describes whether the approach implements a method to automatically detect crime hotspots, which is a crucial issue for the accuracy and the effectiveness of the whole crime forecasting task. The proposed algorithm and the approaches presented in [4,9] implement methods to detect crime hotspots from raw crime data, whereas the rest of the related works rely on pre-defined regions, like Community Areas [10], Prescincts [13], city cells [11]. The limit

Table 1
Comparison of several approaches proposed in literature.

	Crime Hotspots Detection	Hotspot Detection Approach	Crime Hotspot Shapes	Crime Predictors Approach
The proposed approach	*Yes*	*Density-based clustering*	*any shape*	*ARIMA*
Ref. [4]	Yes (ActivitySpace)	Probabilistic framework	any shape	Not Available (only location prediction)
Ref. [10]	No (Comm. Area)	No (predefined area)	Comm. Area shape	Negative Binomial Regression (NBR)
Ref. [13]	No (Precincts)	No (predefined area)	Precinct shape	Holt Exponential Smoothing (HES)
Ref. [9]	Yes	Support Vector Machine	any shape	Not Available (only location prediction)
Ref. [11]	No (grid cells)	equal sized grid cell	square	Recurrent Neural Networks (RNN)

of the latter approaches is that they rely on a static subdivision of regions and categorization of them as region of crime interests, which could lead to regions not interesting in terms of crime analysis. Differently, our approach and the works in [4,9] are able to identify data-driven relevant locations, instead of being statically defined a priori.

Crime hotspot detection approach. We also classified the compared systems on the basis of the approach used to detect crime dense regions, when applicable. The approaches proposed in [4,9] exploit a probabilistic framework and Support Vector Machine (SVM) approaches, opportunely adapted to deal with crime data. On the other side, the approaches presented in [10,11,13], as previously highlighted, use no detection approach as they rely on pre-defined regions.

Crime hotspot shapes. Another important feature for classification purpose is the shape of the crime hotspots. In fact, this feature allows to assess the ability of the detection approach in identifying any possible dense spatial area, regardless of the shape. The more shapes the algorithm is able to catch, the better the accuracy and effectiveness of the detected dense regions. Our approach and the work described in [4,9] are able to detect regions of any shape (e.g., circular, rectangular, linear) while the other works [10,11,13] deal with only specific region shapes.

Crime Predictor approach. This feature classifies the systems on the basis of the approach used to detect crime predictors. Specifically, our approach exploits ARIMA models, while the approaches presented in [10,11,13] use Negative Binomial Regression, HES and RNN models, respectively. Differently from ours, the other approaches [4,9] perform only crime location prediction and they do not consider crime trend analysis.

3. Problem definition and goal

We begin by fixing a proper notation to be used throughout the paper. Let $T = \langle t_1, t_2, \ldots, t_H \rangle$ be an ordered timestamp list, such that $t_h < t_{h+1}, \forall_{0<h<H}$, and where all t_h are at equal time intervals (e.g., every hour, day, week, or year). Let $\mathcal{D}$ be a dataset collecting crime instances, $\mathcal{D} = \langle D_1, D_2, \ldots, D_N \rangle$, where each D_i is a data tuple described by the following features: *latitude* and *longitude* (coordinates of the places the crime occurs), t (time the crime happens at, with $t \in T$), *type* (the crime typology, i.e. robbery, theft, assault, etc.). Now, let us consider a future temporal horizon, $S = \langle t_w, t_{w+1}, \ldots \rangle$, with $w > H$. The goal of the analysis is to find models for reliably predicting the number and location of crimes at a given timestamp $t_w \in S$. More specifically, our analysis aims at achieving the following goals:

1. discover a set $\mathcal{CDR}$ of *crime dense regions* (*blobs* or *hotspots*), $\mathcal{CDR} = \{CDR_1, \ldots, CDR_K\}$, where a *crime dense region* CDR_k is a spatial area which criminal events occur in with an higher density than other areas in the city;
2. extract a function $F_{ncrime} : \mathcal{S} \rightarrow (\mathcal{CDR}, \mathcal{R})$, that given a timestamp $t_w \in S$ states the number of crimes $N \in \mathcal{R}$ that are predicted to happen in each *crime dense region* $CDR_i \in \mathcal{CDR}$ at the timestamp t_w.

4. The proposed approach

This section describes the algorithm that we have designed to discover *spatio-temporal predictive models* from crime data. Specifically, Section 4.1 depicts the main steps of the proposed approach and its meta-code, whereas Sections 4.2 and 4.3 describe in details the procedures for crime dense regions detection and crime predictors extraction.

4.1. The algorithm

Fig. 1 sketches the general idea of the algorithm through a graphic representation of the whole process as a sequence of three main steps. The input data of the analysis is the set of collected crime data to be processed. The first step of the algorithm consists in the *detection of crime dense regions* from the original dataset. The goal of this step is detecting areas (i.e., polygons, blobs) where crime events occur with greater density than other adjacent areas, automatically traced by the algorithm without any pre-fixed division in areas. This task can be modeled as a geo-spatial clustering instance and can be solved, as described below, using clustering algorithms that process both spatial and temporal crime data. The final result of this step consists of K clusters, where each cluster corresponds to a crime dense region. The number of detected regions (i.e., number of clusters) can be fixed a-priori or automatically detected, depending on the specific clustering algorithm. The second step consists in the *spatial data splitting* of the original crime data, based on the clustering model discovered at the previous step. In other words, the points of the original crime data events assigned to the ith cluster are transformed in a

Fig. 1. Spatio-Temporal Crime Prediction Steps.

STCP($\mathcal{D}$)
Input:
$\mathcal{D}$: crime dataset;
Output:
$\mathcal{CDR} = \{CDR_1, \ldots, CDR_K\}$: a set of K crime dense regions;
$\mathcal{F} = \{F_1, \ldots, F_K\}$: a set of K crime predictors;

L1: $\{CDR_1, \ldots, CDR_K\} \leftarrow$ DiscoverCrimeRegions($\mathcal{D}$)
L2: $\{D_1^{\perp}, \ldots, D_K^{\perp}\} \leftarrow$ BuildCrimeTSData($\mathcal{D}, CDR_1, \ldots, CDR_K$)
L3: **for** each $k = 1, ..., K$ **do**
L4: $F_i \leftarrow$ DiscoverLocalCrimePredictor($D_i^{\perp}$)
L5: $\mathcal{F} \leftarrow \mathcal{F} \cup F_i$
L6: **end for**
return $(\mathcal{CDR}, \mathcal{F})$

Fig. 2. Spatio-Temporal Crime Prediction Algorithm.

time series and gathered in the ith output dataset, for $i = 1, \ldots, K$. At the end of this step, K different time series datasets are available, each one containing the time series of crimes occurred in its associated dense region. The third step is aimed at *extracting a specific crime prediction model for each crime dense region* (or the most representative regions), analyzing the crime data split during the previous step.

The meta-code of the Spatio-Temporal Crime Prediction algorithm (STCP) is reported in Fig. 2. The algorithm receives in input $\mathcal{D}$, i.e. the crime dataset, and returns the discovered knowledge models, i.e., the crime dense region set $\mathcal{CDR} = \{CDR_1, \ldots, CDR_K\}$ and the crime predictor set $\mathcal{F} = \{F_1, \ldots, F_K\}$. It is worth noting that this meta-code is parametric with respect to the algorithm for crime dense region detection and crime predictors, and we will give additional details (about the specific algorithms exploited in this work) in the following two sub-sections. The algorithm begins by performing a spatial clustering task over the dataset $\mathcal{D}$, aimed at detecting dense regions (i.e., hot spots) of crimes. This is performed by the DiscoverCrimeRegions() method, which extracts K spatial clusters, each one representing a detected dense region of crimes (line L1). As soon as this step is completed, the crime dataset is transformed in K time-series datasets on the basis of the discovered clustering model extracted at the previous step. Specifically, this task is executed by the BuildCrimeTSData() method (line L2), which processes the original dataset $\mathcal{D}$ and transforms it in the time series dataset collection $\mathcal{D}^{\perp} = \{D_1^{\perp}, \ldots, D_K^{\perp}\}$, where each $D_i^{\perp}$ is the time series of crimes geo-localized in the area $CDR_i \in \mathcal{CDR}$ (detected during the previous step). At the end of this step, K different time-series datasets are available. Finally, for each $D_i^{\perp}$, the DiscoverLocalCrimePredictor() method discovers a predictive model (lines L3–L6) to forecast the number of crimes that will happen in the specific area CDR_i (associated to $D_i^{\perp}$). The whole model returned by the algorithm, comprising the crime dense region set $\mathcal{CDR} = \{CDR_1, \ldots, CDR_K\}$ and the crime predictor set $\mathcal{F} = \{F_1, \ldots, F_K\}$, can be used for spatio-temporal crime forecasting.

4.2. Detection of crime dense regions

The DiscoverCrimeRegions() method (line L1) performs a spatial clustering of the dataset, where each cluster represents a dense region of crimes. The density-based notion is a common approach for clustering, whose inspiring idea is that objects forming a dense region should be grouped together into one cluster. In our implementation, this step is performed by applying DBSCAN [15], a popular density-based clustering algorithm that finds clusters starting from the estimated density distribution of the considered data. We have chosen the DBSCAN algorithm because it has the ability to discover clusters with arbitrary shape such as linear, concave, oval, etc. and (in contrast to other clustering algorithms proposed in literature) it does not require the predetermination of the number of clusters to be discovered. Basically, the algorithm finds clusters with respect to the notion of density reachability among points: a point is directly density-reachable from another point if it is not farther away than a given distance (ϵ) (i.e., is part of its neighborhood) and if it is surrounded by sufficiently many points (*minPts*). In the considered context, a cluster corresponds to a crime dense region. Moreover, to capture the

dynamic changing of clusters, we compute the density of each data point by weighting it through a *decay factor* which gives less importance to historical information and more weight to recent data: for each data record C_i, we assign it a density coefficient which decreases with as C_i ages: if C_i occurs at the timestamp t_i, its density coefficient is weighted by $\lambda^{t_{max}-t_i}$, where $\lambda \in (0, 1)$ is a constant called the decay factor, and t_{max} is the most recent timestamp. Finally, DBSCAN requires the user to specify the radius of the neighborhood (i.e., ε) and the minimum number of objects it should have (i.e., *minPoints*), whose values affect size and density of the discovered clusters. Generally, an optimal setting of its parameters is complex to be achieved and requires specific techniques; nevertheless, such a topic is out of the scope of this paper.

4.3. Extraction of crime predictors

Given a specific crime dense region (or crime hotspot), the DiscoverLocalCrimePredictor() method (line L4 in Fig. 2) discovers a predictive model to forecast the number of crimes that will happen in its specific area. In our implementation, this has been performed by the *Seasonal AutoRegressive Integrated Moving Average* model (Seasonal ARIMA, or SARIMA [16]), which is defined as a combination of auto-regression, moving average and difference modeling. Briefly, having the time series $\{y_t : t = 1...n\}$, where y_t is the value of the time series at the timestamp t, an $ARIMA(p, d, q)$ model is written in the form

$$y_t^{(d)} = c + \phi_1 y_{t-1}^{(d)} + \cdots + \phi_p y_{t-p}^{(d)} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

where c is a correcting factor, $\phi_1, \ldots, \phi_p$ are the regression coefficients of the auto-regressive part, $\theta_1, \ldots, \theta_q$ are the regression coefficient of the moving average part, $y_{t-1}, \ldots, y_{t-p}$, $e_{t-1}, \ldots, e_{t-q}$ are lagged values of y_t and lagged errors ($p+q$ predictors), and e_t is white noise and takes into account the forecast error. In our study we exploit *seasonal ARIMA* models, which are an extension of the classic ARIMA. A seasonal ARIMA model is formed by including additional seasonal terms in the classic ARIMA models previously introduced. The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model. In the final formula, the additional seasonal terms are simply multiplied with the non-seasonal terms. A seasonal ARIMA model is defined as $ARIMA(p, d, q)(P, D, Q)_m$, where m is a periodicity factor, (p, d, q) and (P, D, Q) are the orders of the auto-regressive, differencing and the moving average part for the non-seasonal and seasonal model, respectively [16].

5. Analysis and experimental results

To evaluate the performance and the effectiveness of the approach described above, we carried out an extensive experimental analysis by executing different tests in two real-world case-studies, i.e., two large areas of Chicago (CHI) and New York City (NYC). For each city, the goal of our analysis comprises detecting the most significant crime dense regions and discovering effective predictive models, which can estimate the number of crimes that are likely to happen in the future. We also performed a comparative analysis of our results with respect to other algorithms proposed in literature. The rest of this section is organized as follows. Section 5.1 presents the analysis and the most important results carried out on the Chicago data. Section 5.2 describes the results obtained by analyzing the New York City data. Section 5.3 compares the results obtained by exploiting autoregressive models with other approaches used for regression analysis.

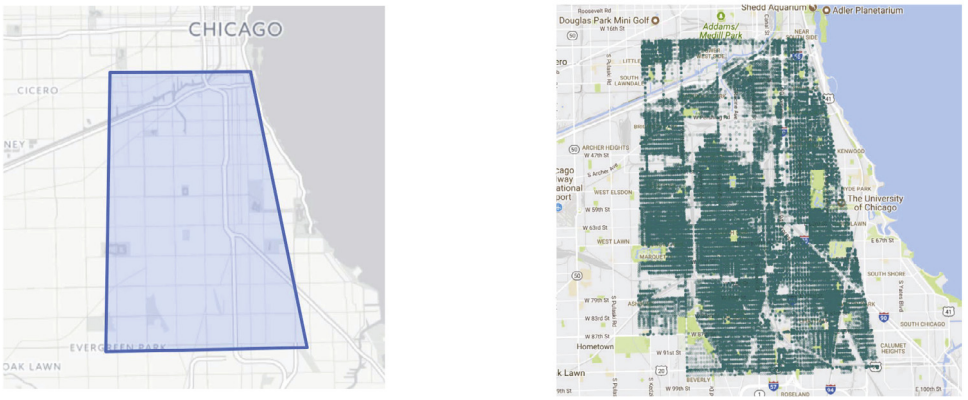
5.1. Chicago: Experimental results

In the following sub-sections we describe the main steps of our analysis carried out on Chicago data: (*i*) data description and gathering, (*ii*) crime dense region detection, (*iii*) training and evaluation of the regressive models.

5.1.1. Data description

The data that we used to train the models and perform the experimental evaluation is housed on Plenario, a publicly available data search and exploration platform that was developed (and currently managed) by the University of Chicago's Urban Center for Computation and Data. Crime data has been gathered from the *'Crimes - 2001 to present'* dataset, a real-life collection of instances describing criminal events occurred in Chicago from 2001 to present (the repository is updated every week, so it is kept up-to-date minus the most recent seven days). As a pilot research study, in this work we focus our experiments on a large area of Chicago, which is shown in Figs. 3(a) and 3(b). The selected area includes different zones of the city, some growing in terms of population, others in terms of business activities, with different crime-densities over their territory (so making it interesting for crime analysis). Its perimeter is about 52 KM and its area is approximately 135 KM^2. Starting from the *'Crimes - 2001 to present'* dataset, we collected all crime events within the bounded area over 16 yr (834 weeks), from January 2001 to December 2016. The total number of collected crimes is 1,897,682, while the average number of crimes per week is 2275. The total size of this dataset is 418 MB.

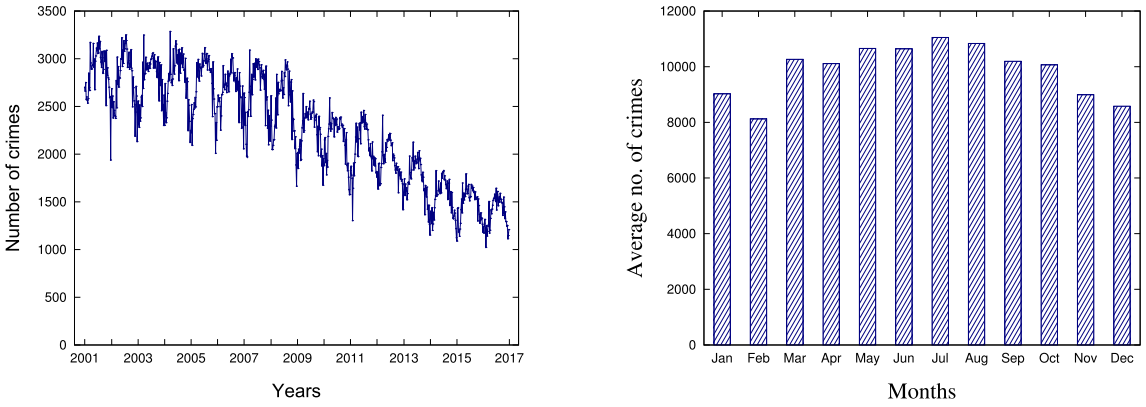
Figs. 4(a) and 4(b) show a preliminary view of the collected crime data, which provides some hints about data trends and distribution. Fig. 4(a) reports the time plot of the observed crime data, in which the number of crimes is plotted versus the time of observation. The plot immediately reveals some interesting features. First, it is evident that the number of crimes is decreasing over the time period, showing a general clear *decreasing trend* in the data. Second, a repeating *seasonal pattern* within each year is clearly observable, that seems to decrease in size (magnitude) as the overall crime counts in the series



(a) Polygon of the area traced in Plenar.io.

(b) Geo-localized crime events.

Fig. 3. Selected area of Chicago and geolocalized crime events (2001–2016).



(a) N. of crimes vs time

(b) Distribution by month

Fig. 4. CHI crime data: number of crimes vs time and their distribution by month.

decreases. From the plot, we see that the occurrence of crimes typically increases in late Spring, peaks during the Summer, decreases in Autumn, and generally dips in Winter. A clearer view of the seasonality hidden in the data can be seen in Fig. 4(b), which shows the distribution of the average number of crimes by month. The histogram shows that the number of crimes in the city area under observation varies significantly between different periods of the year. In particular, the number of criminal events is highest in July (with 11,050 crimes on average), and lowest in February (with 8124 crimes, on average).

To perform the regression task and its validation, we split the original dataset in two partitions: the training set and the test set. The first is used to discover the relationships inside data while the second is used for evaluating whether the discovered relationships hold generally. In our case, the overall crime dataset has been split with respect to the number of years: the training set contains the crime data of the first 13 yr (2001–2013, 678 weeks), while the test set holds the crime data of the last 3 yr (2014–2016, 156 weeks). As described in the following sub-sections, we trained the knowledge model (i.e., crime dense regions and crime predictors) using data from January 2001 to December 2013 and we used the trained model to forecast the crime events from January 2014 to December 2016, to assess the quality of the predictions.

5.1.2. Detection of crime dense regions from the training set

As described in Section 4.2, crime dense regions are detected by applying an our ad-hoc modified version of DBSCAN, which exploits a decay factor that gives a higher weight to recent crime events. Moreover, in order to detect high quality crime dense regions, it is necessary to tune the key parameters of the algorithm so as to improve results' performance. In particular, the values of the DBSCAN's parameters ϵ and *minPts* determines the size of the clusters, as they represents the minimum crime density required by an area to be part of a cluster. On the one hand, the bigger ϵ, the larger is the

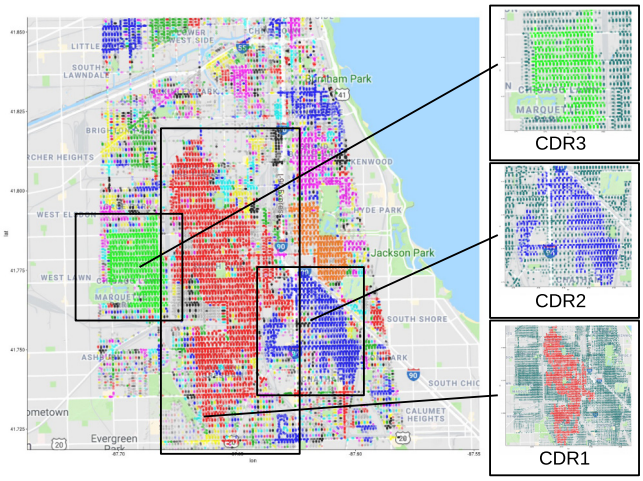


Fig. 5. Detected crime dense regions in the selected area of Chicago. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Extension of the wider Crime Dense Regions w.r.t. the whole area considered.

Region	Extension (KM^2)	Extension (%)	Crimes (#)	Crimes (%)
Whole Area	135.02	100.00%	1,896,782	100%
Crime Dense Region 1	17.04	12.62%	461,996	24.35%
Crime Dense Region 2	6.03	4.47%	168,159	8.86%
Crime Dense Region 3	4.59	3.40%	131,154	6.91%

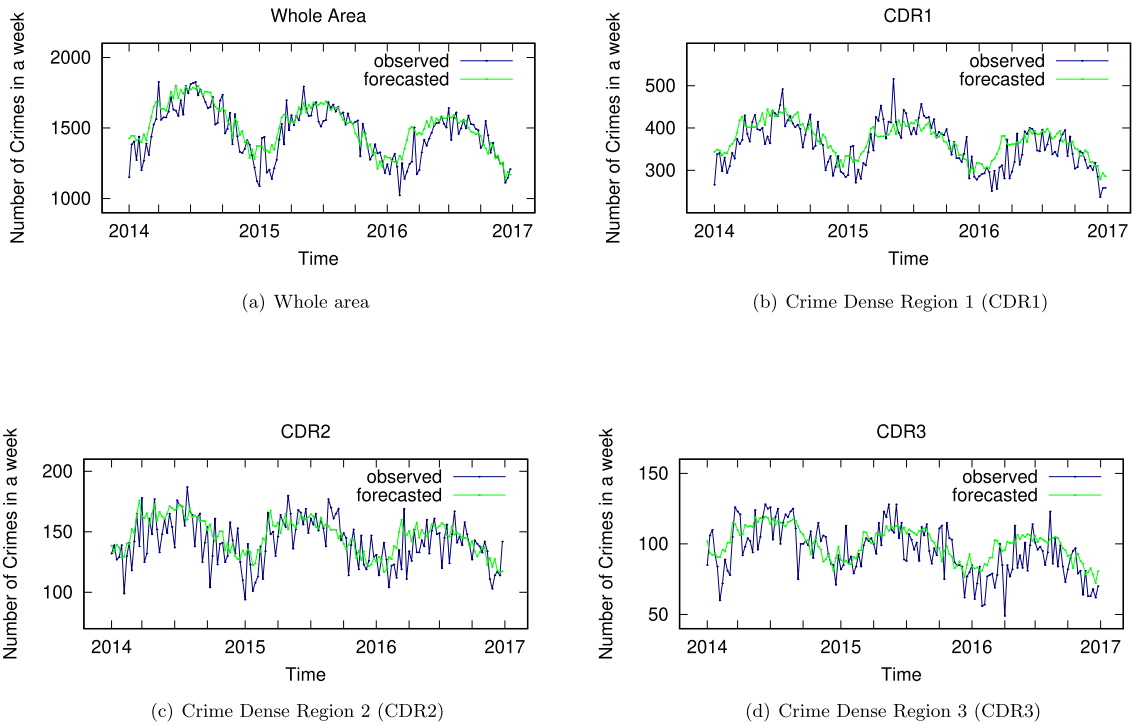
extension of the dense regions detected: this results in the discovery of large regions that actually are no longer dense. On the other hand, the smaller ϵ, the smaller the cluster sizes, resulting in a high number of dense regions detected that could be (because of their small sizes) not significant for the analysis. Conversely, growing the *minPts* value results in increasing the fragmentation of the clustering assignment produced. The values of ϵ and *minPts* are, thus, critical to the accuracy of the dense region detection phase and for the right balance among separability, compactness and significance of clusters. We present here the results achieved by fixing $\epsilon = 83.25$ m and $minPts = 20$, which have been assessed through several experimental tests and best suits our application scenario and the considered dataset.

Crime dense regions discovered through our analysis are shown in Fig. 5, where each region is represented by a different color. Interestingly, this image shows how crime events are clustered on the basis of a density criteria; for example, the algorithm detects eight significant crime regions clearly recognizable through different colors: a large crime region (in red) in the central part of the area along with seven smaller areas (in green, blue and light-blue) on the left and right side, corresponding to zones with the highest concentration of crimes. The three largest crime dense regions (*CDR1*, *CDR2*, and *CDR3*) are zoomed-in on the left side of Fig. 5. Many other smaller regions representing very local high-density crime zones are distributed in the whole area. Table 2 shows the extension of the three largest crime dense regions (*CDR1*, *CDR2*, and *CDR3*), with respect to the whole area. Overall, these regions cover about the 20.5% of the whole area extension, and about the 40% of the crime events detected in the whole area between 2001 and 2016.

5.1.3. Training and evaluating the regressive crime models

As described in Section 4, the next steps of the algorithm consist of (*i*) the *spatial data splitting* of the original crime dataset (aimed at building a time series for each discovered dense region), and (*ii*) the training of local crime predictors (as ARIMA models) for each dense region. Specifically, considering the three largest crime dense regions detected by our algorithm, the auto-regressive models trained from the input data were $ARIMA(1, 1, 2)(1, 1, 2)_{52}$, $ARIMA(4, 1, 3)(0, 1, 1)_{52}$ and $ARIMA(1, 1, 1)(0, 1, 1)_{52}$ for the first, second and third region, respectively. It is worth noting that the predictive crime models differ among regions, showing that each area presents specific crime trends and patterns.

In order to assess the effectiveness and accuracy of the regressive functions, we performed an evaluation analysis on the test set consisting of the last three years of data (i.e., years 2014–2016). In particular, for each crime dense region and for the whole area, its respective ARIMA model has been used to predict future values of the number of crimes that are likely to happen in that region, week by week. The prediction of the type of crimes is beyond the scope of this work and it will be studied in a further research activity. Fig. 6 shows observed and forecasted data (plotted in blue and green, respectively) for the test set period. We note that forecasted data adhere very well to the observed data over the whole test set period.



(a) Whole area

(b) Crime Dense Region 1 (CDR1)

(c) Crime Dense Region 2 (CDR2)

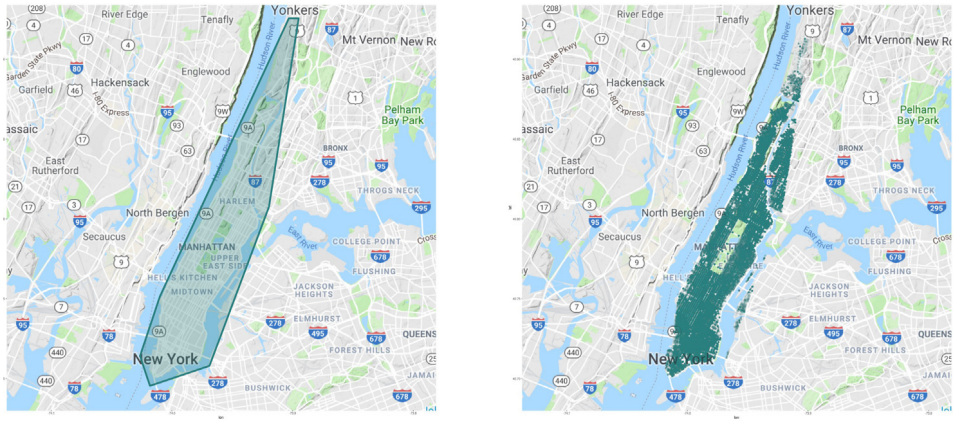
(d) Crime Dense Region 3 (CDR3)

Fig. 6. Number of crimes observed and forecasted (blue and green lines) on the Chicago test set, for the whole area and the top 3 largest crime dense regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Forecast error measures vs years, for the whole area and the top three largest crime dense regions in Chicago.

Time	MAE				MAPE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	88.86	30.20	14.47	11.15	6.19	8.68	10.86	11.90
2015	74.54	28.24	12.13	9.24	5.42	7.60	8.94	9.62
2016	81.47	31.04	12.86	13.83	6.29	10.14	9.99	18.66
Time	ME				RMSE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
hline 2014	−62.96	30.19	−8.36	−2.67	108.34	35.98	19.01	13.57
2015	−27.05	4.75	−1.80	−1.57	97.77	34.61	14.98	11.43
2016	−48.77	−24.94	−5.36	−11.02	115.16	38.31	15.95	16.66

Now, let us give a quantitative evaluation about the performance of the regressive models and their effectiveness in making predictions on the corresponding test sets. To this end, we computed four error measures (*MAE*, *MAPE*, *ME*, *RMSE*), which are commonly used in regressive analysis literature to quantify forecast performance. Table 3 reports the values of the four error measures described above for the whole area and the three largest crime dense regions, by considering one-year-ahead, two-year-ahead and three-year-ahead prediction horizons. Looking at the values in the table, we can observe the MAE decreases when the areas of regions are smaller and smaller. For example, considering one-year-ahead forecasting, the MAE monotonously decreases from 88.86 (whole area) to 30.20, 14.47 and 11.15 (three crime dense regions, ordered by decreasing size), and similarly all other years. This is a reasonable result, because predictions appear more precise both in terms of specific identification of the areas and in terms of forecasting accuracy, thus giving a more detailed information to city administrator and police officers for planning how to distribute resources and efforts in the different regions of the city. Finally, considering the MAPE, we observe that percentage errors are very low as well. In fact, the table shows that the maximum MAPE forecasting error ranges from 8.68% to 11.90% for the first year, from 7.60% to 9.62% for the second year, and from 10.14% to 18.66% for the third year, which represents a very interesting result. To the best of our knowledge, these results exceed those of other approaches proposed in the crime forecasting literature. As a final consideration, we observe from Fig. 6 that the regressive models for each area tend to slightly over-forecast the number of crimes with respect to those that actually occurred.



(a) Polygon of the area. (b) Geo-localized crimes.

Fig. 7. Selected area of New York City and geolocalized crime events (2001–2016).

5.2. New York City: Experimental results

This section presents the analysis performed on New York City crime data. As with the previous case study, the main steps are described in four subsections: (*i*) data description and gathering, (*ii*) crime dense region detection, and (*iii*) training and evaluation of the regressive models.

5.2.1. Data description

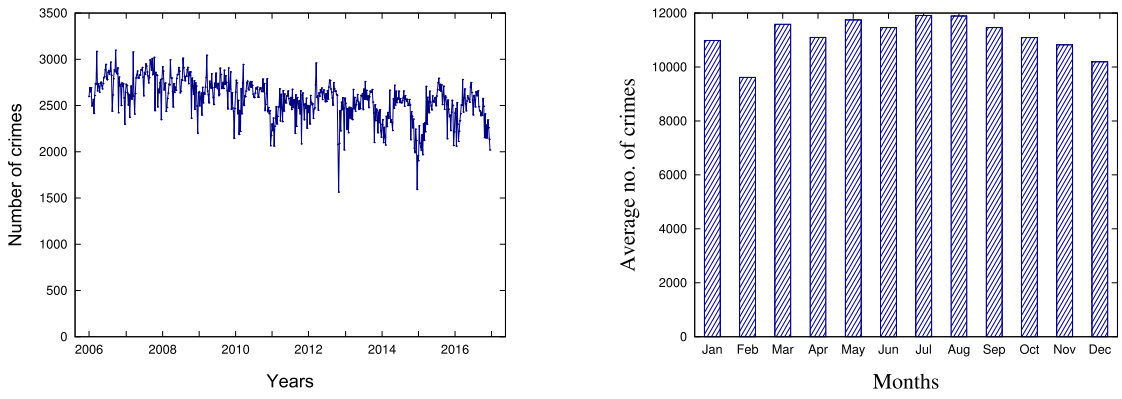
The data that we used to train the models and perform the experimental evaluation for New York City is housed on NYC Opendata [7], a publicly available resource managed by the Mayor's Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT). We focus our experiments on the Manhattan borough of New York City, which is shown in Figs. 7(a) and 7(b). The selected area represents one of the most well-known and dense urban areas in the world, growing in terms of population, business activities, and mobility patterns. Its perimeter is about 44 KM and its area is roughly 76 KM^2. Starting from the *'NYPD Complaint Data Historic'* dataset, we collected all crime events that were recorded within the bounded area in 11 yr (572 weeks), from January 2006 to December 2016. The total number of collected crimes is 1,472,305, and the average number of crimes per week is 2573. The total data size is about 1.3 GB.

New York City data trends and distribution are shown in Figs. 8(a) and 8(b). As a preliminary view, we observe that there are several differences with the distribution of the Chicago data. Fig. 8(a) reports the time plot of the observed crime data, in which the number of crimes are plotted versus the time of observation. In contrast to the Chicago data (showing steady decrease in total crimes over time), the chart clearly shows that the number of crimes exhibits a *stable trend* until the year 2010, followed by a *smooth decreasing trend* from 2010 to 2012, and a *stable trend* again from 2012. Second, a yearly seasonal pattern is observable, which increases in size (magnitude) and becomes more evident from the year 2012. In general, we can infer that the occurrences of crimes usually achieve peaks during both the Spring and the Summer (differing from the Chicago data), decrease in Autumn and generally have dips in Winter. The seasonality component can be observed in Fig. 8(b), which shows the distribution of the average number of crimes by month. The average number of criminal events is highest during the Summer (11,910 in July and 11,188 in August), but there are also some peaks in the Spring (11,586 in March and 11,742 in May). The lowest count is in February (with 9,615 crimes on average). Data splitting into training set and test set has been performed as follows: the training set contains the crime data of the first 8 yr (2006–2013; 416 weeks), while the test set holds the crime data of the last 3 yr (2014–2016; 156 weeks). The crime dense regions and crime predictors discovered on the training set, as well as a discussion about the quality assessment of the models on the test set, are described in the following two subsections.

5.2.2. Detection of crime dense regions from the training set

As discussed in Section 4.2, the optimal parameter setting of DBSCAN to discover crime dense regions is not an easy task. The goal is to detect a suitable tuning of ϵ and *minPts* values, which are key factors for the accuracy of the dense region detection phase and for the right balance among separability, density, compactness and significance of clusters. We present here the results achieved by fixing $\epsilon = 120.36$ m and $minPts = 20$, which have been assessed through several experimental tests to best suit our application scenario and the considered dataset.

Crime dense regions of New York City, discovered through the DBSCAN algorithm, are shown in Fig. 9. The algorithm detects seven significant crime regions clearly recognizable through different colors: a large crime region (in red) covering



(a) N. of crimes vs time (b) Distribution by month

Fig. 8. NYC crime data: number of crimes vs time and their distribution by month.

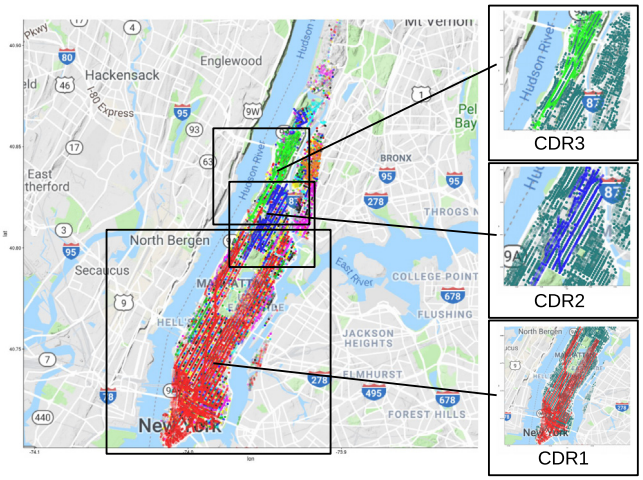


Fig. 9. Detected crime dense regions in the selected area of New York City. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

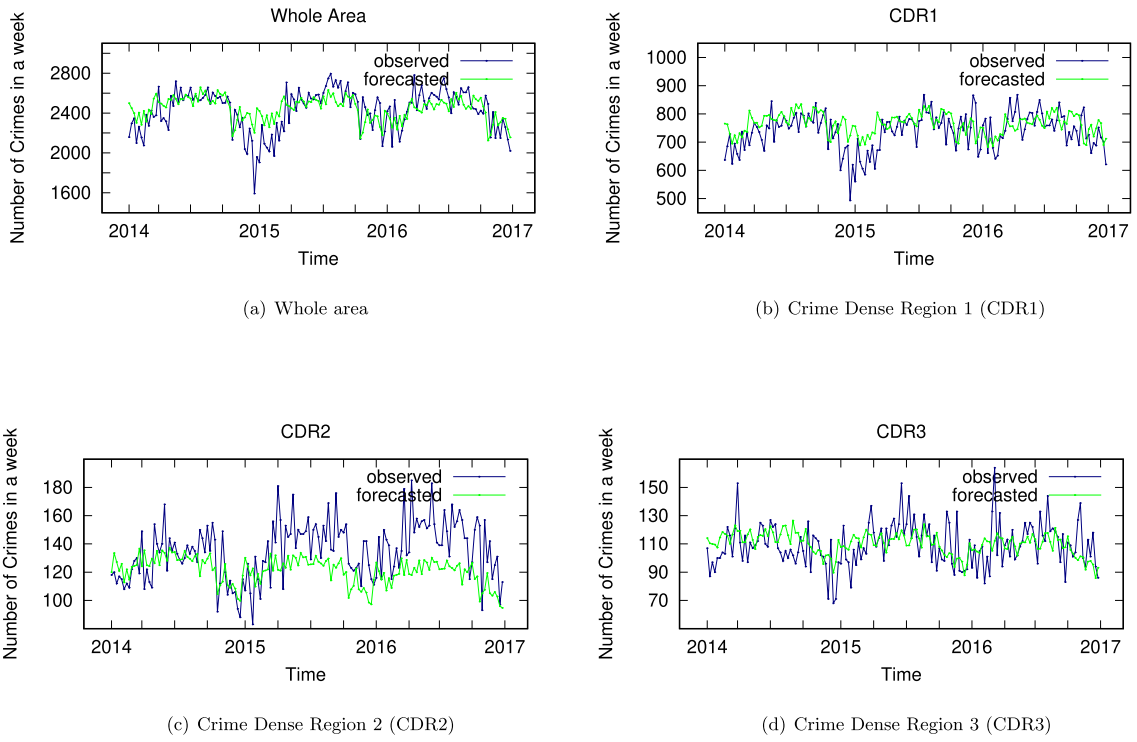
Table 4
Extension of the wider Crime Dense Regions w.r.t. the whole area considered.

Region	Extension (KM^2)	Extension (%)	Crimes (#)	Crimes (%)
Whole Area	76.41	100.00%	1,472,305	100%
Crime Dense Region 1	13.57	17.77%	452,113	30.70%
Crime Dense Region 2	2.33	3.05%	79,803	5.42%
Crime Dense Region 3	2,71	3.55%	69,268	4.70%

Midtown and Lower Manhattan (including the Financial District), and other six smaller areas (in green, purple, blue and light-blue) on the upper East and West sides, corresponding to zones with the highest concentration of crimes. The three largest crime dense regions (*CDR1*, *CDR2*, and *CDR3*) are zoomed-in on the left side of Fig. 9. We observe that there are many other smaller regions representing very local high-density crime zones, whose small size make them less interesting for our analysis. Table 4 shows the extension of the three largest crime dense regions (*CDR1*, *CDR2*, and *CDR3*), with respect to the whole area. Overall, these regions cover about 24.5% of the whole area, and about 40% of the crime events detected in the whole area between 2006 and 2016.

5.2.3. Training and evaluating the regressive crime models

Crime regressive models have been extracted for the three largest crime dense regions detected by our algorithm for Manhattan. Specifically, the auto-regressive models trained from the input data were $ARIMA(1, 1, 2)(0, 1, 1)_{52}$, $ARIMA(1, 1, 1)(0, 1, 1)_{52}$ and $ARIMA(2, 1, 3)(0, 1, 1)_{52}$ for the first, second and third regions, respectively. It is worth noting



(a) Whole area

(b) Crime Dense Region 1 (CDR1)

(c) Crime Dense Region 2 (CDR2)

(d) Crime Dense Region 3 (CDR3)

Fig. 10. Number of crimes observed and forecasted (blue and green lines) on the New York test set, for the whole area and the top 3 largest crime dense regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that predictive crime models differ among regions (as also seen for the Chicago case study), as evidence that each area presents specific crime trends and patterns.

The evaluation of the regressive functions trained on the New York City data has been performed on the test set, which consists of the last three years of data (i.e., years 2014–2016). Fig. 10 shows observed and forecasted data (plotted in blue and green, respectively) for the test set period. It is worth noting that forecasted data adhere very well to the observed data over the whole test set period. Only in the Crime Dense Region 2 case, we observe that there is an evident difference among the two curves, and in particular how the forecasting trend assumes lower values than the real trend.

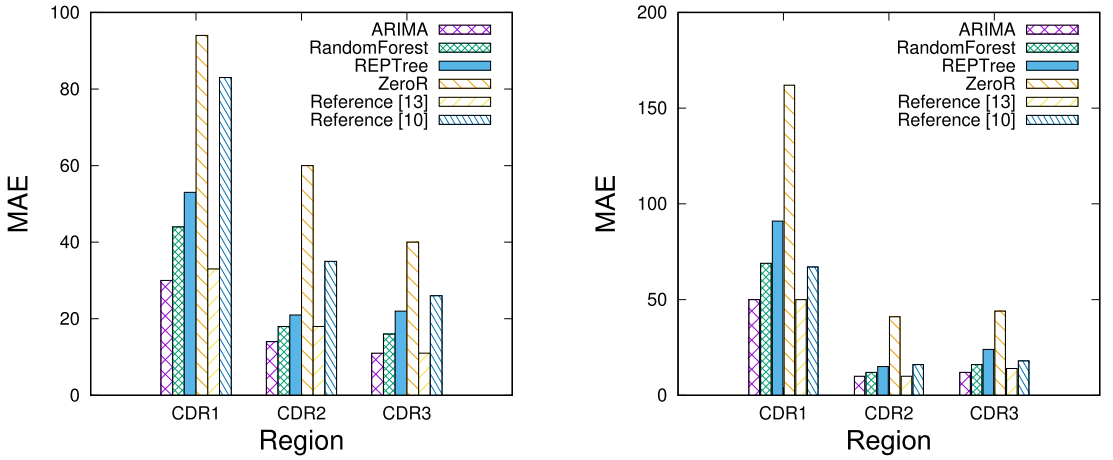
Now, let us give a quantitative evaluation about the performance of the regressive models and their effectiveness in making predictions on the corresponding test sets. To this end, forecasting performance have been evaluated using the *MAE*, *MAPE*, *ME* and *RMSE* error measures, for several time horizons. The values of the four error measures are reported in Table 5, for the whole area and the top three largest crime dense regions, by considering one-year-ahead, two-year-ahead and three-year-ahead prediction horizons. We can observe that MAE values decrease when the areas of regions are smaller. For example, considering one-year-ahead forecasting, the MAE decreases from 135.30 (whole area) to 52.15, 10.56 and 12.46 (three crime dense regions, ordered by decreasing sizes), and similarly all other years. A similar trend has been observed also for the Chicago case study, and it is confirmed for the NYC case study. It is worth noting that it is a reasonable result, because forecasts appear more precise both in terms of specific identification of the areas and in terms of forecasting accuracy. As a final consideration, we observe from Fig. 10 that the regressive models for the whole area, Crime Dense Region 1 and 3 (Figs. 10(b) and 10(d)) adhere to the observed data much better than the Crime Dense Region 2 case (Fig. 10(c)).

5.3. Comparative analysis with other approaches

To make our evaluation more accurate and complete, we performed a comparative analysis of several approaches for crime predictors extraction. Specifically, we evaluated the performance of ARIMA models versus three classic regression algorithms (i.e., *RandomForest* [17], *REPTree* [18], *ZeroR* [19]) and versus the approaches [10] and [13] specifically proposed in the crime forecasting literature. To perform the comparative analysis, we evaluated the forecasting performance of the different approaches on the test set of the three areas (both in Chicago and New York City), versus different prediction horizons. The results for the algorithms were obtained by performing an accurate tuning of the input parameters: for each dataset, different runs were executed for different values of the parameters, then the best results were selected. The results shown below only refer to the run with the best combination of parameters. Fig. 11 summarizes the results of the comparison, showing the achieved Mean Absolute Error (*MAE*) for one-year-ahead forecasts, for the three highest crime dense regions

Table 5
MAE, MAPE, ME and RMSE prediction errors vs years, for the whole area and the top three largest crime dense regions in New York City.

Time	MAE				MAPE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	135.30	52.15	10.56	12.46	6.17	7.26	8.15	12.59
2015	141.68	51.51	20.86	11.06	6.04	7.42	14.95	9.87
2016	117.49	47.85	25.05	11.80	4.79	6.41	16.64	10.58
Time	ME				RMSE			
	Area	CDR1	CDR2	CDR3	Area	CDR1	CDR2	CDR3
2014	−80.05	−41.17	2.11	−6.55	184.81	70.99	14.23	14.73
2015	−0.6	−28.57	15.63	1.51	177.08	66.08	24.60	14.45
2016	31.21	−8.00	23.84	2.68	143.73	57.80	29.85	16.19



(a) Chicago: MAE (b) New York City: MAE

Fig. 11. Comparative analysis among several approaches, evaluating the Mean Absolute Error (*MAE*) of the crime dense regions, for CHI (a) and NYC (b).

in Chicago and New York City. In particular, we can see that the ARIMA approach generally achieves greater accuracy than other algorithms. In fact, considering both the Chicago and New York datasets, the ARIMA models perform better for all of the highest crime-dense regions. Indeed, the performance difference is more evident for larger areas. These results confirm the appropriateness of the autoregressive model and its good performance in the crime prediction domain.

6. Conclusion

This paper presented a general algorithm for Spatio-Temporal Crime Prediction in urban areas, implemented in context of partitioning large areas of cities into sub-areas by detecting crime dense regions (of arbitrary shapes). Such regions are then analyzed and a different forecasting auto-regressive model is tailored specifically for each detected region. Experimental evaluation, performed on two datasets, related to the crime data of wide areas of Chicago and New York City, showed that the proposed methodology can forecast the number of crimes with high accuracy. Furthermore, the approach gives fine-grained information about where crime events are expected to occur. We also presented a comparative analysis with other regressive algorithms, showing that (at the best of our knowledge) the achieved results outperform those of other approaches proposed in the crime forecasting literature so far. In future work, other research issues may be investigated. First, we may further explore the application of other spatial analysis approaches for the detection of crime dense regions, to forecast crime trends on such regions. Specifically, we are interested in studying the application of hierarchical spatial algorithms, which can achieve further splitting of clusters when their sizes are too large. Second, we will correlate the trend of crimes and other events of the city to understand relationships among those, as well as to explore the use of these spatio-temporal prediction algorithms to predict other kinds of events.

References

[1] United Nations Settlements Programme, the state of the world's cities 2004/2005: Globalization and urban culture. Earthscan, 2004.
[2] Cities: The century of the city, Nature 467 (2010) 900–901.

[3] F. Cicirelli, A. Guerrieri, G. Spezzano, A. Vinci, An edge-based platform for dynamic smart city applications, Future Gener. Comput. Syst. 76 (2017).
[4] M. Tayebi, M. Ester, U. Glasser, P. Brantingham, CRIMETRACER: Activity space based crime location prediction, in: Advances in Social Networks Analysis and Mining, ASONAM, 2014 IEEE/ACM International Conference on, 2014, pp. 472–480.
[5] H. Wang, D. Kifer, C. Graif, Z. Li, Crime rate inference with big data, in: Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016.
[6] C. Catlett, T. Malik, B. Goldstein, J. Giuffrida, Y. Shao, A. Panella, D. Eder, E. van Zanten, R. Mitchum, S. Thaler, I.T. Foster, Plenario: An open data discovery and exploration platform for urban science, IEEE Data Eng. Bull. 37 (4) (2014).
[7] New York city opendata, , 2018-09-30.
[8] C. Catlett, E. Cesario, D. Talia, A. Vinci, A data-driven approach for spatio-temporal crime predictions in smart cities, in: Proceedings of the 2018 IEEE International Conference on Smart Computing, SMARTCOMP'18, 2018, pp. 17–24.
[9] K. Kianmehr, R. Alhajj, Crime hot-spots prediction using support vector machine, in: Computer Systems and Applications, 2006. IEEE International Conference on, 2006, pp. 952–959.
[10] H. Wang, D. Kifer, C. Graif, Z. Li, Crime rate inference with big data, in: Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 635–644.
[11] Y. Zhuang, M. Almeida, M. Morabito, W. Ding, Crime hot spot forecasting: A recurrent model with spatial and temporal information, in: 2017 IEEE International Conference on Big Knowledge, ICBK, 2017, pp. 143–150, http://dx.doi.org/10.1109/ICBK.2017.3.
[12] B. Chandra, M. Gupta, M. Gupta, A multivariate time series clustering approach for crime trends prediction, in: Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, 2008.
[13] W. Gorr, A. Olligschlaeger, Y. Thompson, Short-term forecasting of crime, Int. J. Forecast. 19 (4) (2003) 579–594.
[14] P. Chen, H. Yuan, X. Shu, Forecasting crime using the ARIMA model, in: Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on, vol. 5, 2008, pp. 627–630.
[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, AAAI Press, 1996.
[16] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, OTexts.com, 2014.
[17] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[18] A. Hall Mark, H. Witten Ian, Frank Eibe, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.
[19] S. Nasa C., Evaluation of different classification techniques for WEB data, Int. J. Comput. Appl. 52 (9) (2012) 34–40.