2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Forecasting Crimes using Autoregressive Models

Eugenio Cesario
ICAR-CNR
Rende (CS), ITALY
Email: cesario@icar.cnr.it

Charlie Catlett
University of Chicago
Argonne National Laboratory
Chicago, IL, USA
Email: catlett@anl.gov

Domenico Talia
DIMES - University of Calabria
Rende (CS), ITALY
Email: talia@dimes.unical.it

***Abstract*—As a result of steadily increasing urbanization, by 2030 more than sixty percent of the global population will live in cities. This phenomenon is stimulating significant economic and social transformations, both positive (such as, increased opportunities offered in urban areas) and negative (such as, increased crime and pressures on city budgets). Nevertheless, new technologies are enabling police departments to access growing volumes of crime-related data that can be analyzed to understand patterns and trends. Such knowledge is useful to anticipate criminal activity and thus to optimize public safety resource allocation (officers, patrol routes, etc.) through mathematical techniques to predict crimes. This paper presents an approach, based on auto-regressive models, for reliably forecasting crime trends in urban areas. In particular, the main goal of the work is to design a predictive model to forecast the number of crimes that will happen in rolling time horizons. As a case study, we present the analysis performed on an area of Chicago, using a variety of open data sources available for exploration and examination through the University of Chicagos Plenario platform. Experimental evaluation shows that the proposed methodology predicts the number of crimes with an accuracy of 84% on one-year-ahead forecasts and of 80% on two-year-ahead forecasts.**

I. Introduction

The world is rapidly urbanizing and undergoing the largest wave of urban growth in history. According to a United Nations report urban population is expected to grow from 2.86 billion in 2000 to 4.98 billion in 2030 [1]. This translates to roughly 60% of the global population living in cities by 2030. Much of this urbanization is already bringing huge social, economic and environmental transformations and at the same time presenting challenges in city management issues, like resource planning (water, electricity), traffic, air and water quality, public policy and public safety services.

As cities grow, public safety issues also increase. Indeed it has been shown that larger cities have higher crime rates than smaller communities [2], [3]. In fact, a major concern in many countries is spiking crime rates in urban areas along with the projected growth of urban population [4]. Policy-makers and law enforcement agencies will inevitably face enormous challenges deploying perennially scarce resources even more efficiently to apprehend criminals, disrupt criminal networks and effectively deter crime by investing in crime prevention and reduction strategies in urban areas [5], [6], [2]. A major challenge facing all law-enforcement organizations is accurately and efficiently analyzing the growing volume of crime data as well as the similarly increasing diversity of data sources that may have predictive value related to crime. This is particularly the case in large urban areas, where law enforcement agencies lack the tools and technologies to identify useful patterns in these large volumes of time series data. This presents the opportunity to apply data mining methodologies to extract useful information from this data in order to improve the efficacy of urban policing by enabling police departments to better utilize limited resources. Moreover, advanced analytical tools also have the potential to be integrated with existing planning tools, enabling criminal investigators to explore large databases quickly and efficiently without being trained as data scientists. The objective of these methods is to develop effective strategies that will prevent crime or make investigation efforts more effective.

One of the most challenging issues of municipal police departments is to have accurate (short-term and long-term) crime forecasts in the city. In particular, police precincts and patrol districts would benefit greatly from accurate short-term forecasts of crime within small geographic areas. Such forecasts would make it possible to target patrols to areas with forecasted crime increases, remove and redeploy special details in areas with forecasted crime decreases, schedule training and vacations in low crime periods, etc. In a similar way, long-term forecasts of crime trends are very useful to understand if safety in a city area is likely to improve or decline. Moreover, they help to efficiently deploy police officers and other resources, as well as to plan for hiring or other resource enhancements, across the metropolitan area based on forecasted risk.

This paper presents the design and implementation of an approach, based on auto-regressive models, for reliably forecasting crime trends in an urban area. As case study, we present the analysis of crimes within an area of Chicago. The main goal consists in having a predictive model to forecast the number of crimes that will happen in rolling time horizons. Crime data has been gathered by the Plenario platform [7], a Web framework that provides public access to more than one hundred urban datasets. The results of the experimental evaluation show the effectiveness of the approach. In fact, it performs one-year-ahead and two-years-ahead forecasts with an average accuracy of 84% and 80%, respectively. To the best of our knowledge, these results exceed those of other approaches found in crime forecasting literature.

The rest of the paper is organized as follows. Section II reports the most important approaches in crime data mining literature and the most representative projects in such a research field. Section III summarizes the state-of-art of auto-regressive and moving average models, and their combination to build ARIMA (AutoRegressive Integrated Moving Average) models. Section IV fixes the problem statement and goals of

978-1-5090-4065-0/16 $31.00 © 2016 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.138

our analysis. Section V is the main part of the paper and describes the analysis we performed in an area of Chicago to forecast the number of crimes predicted to happen in the future. It describes the data, the training of the regressive model and its evaluation on real data. Finally, Section VI concludes the paper and outlines avenues for future work.

II. Related Work

The crime data mining field is a relatively new research area, whose goal is to identify patterns in criminal behaviors, in order to anticipate criminal activity and predict/prevent crimes [8]. Several classic data mining techniques have been used for crime analysis, such as association rule mining [2], [9], classification [10], and clustering [5], [11]. In this section we will briefly review some of the most representative research and projects.

A general framework for crime data mining, experimented on some experiences in collaboration with the Tucson and Phoenix Police departments, is presented in [12]. In particular, the paper describes three examples of its use in reality. First, entity extraction algorithms have been used to automatically identify persons, addresses, vehicles, and personal characteristics from police narrative reports (that usually contain many typos, spelling errors, grammatical mistakes, etc.). Second, a text mining algorithm has been experimented with for deceptive-identity detection, to discover the real identity of suspects that have given false names, faked birth dates and/or false addresses. Third, a concept-based approach has been exploited to identify subgroups or key members in criminal networks, to study interaction patterns among them.

CrimeTracer, a random walk-based approach for spatial crime analysis and crime location prediction, has been recently presented in [4]. In particular, the methodology is based on a probabilistic framework to model the spatial behavior of known offenders within areas they are most familiar with, called activity spaces. Experiments carried on real-world crime data have shown that criminals, rather than venture into unknown territory, frequently commit opportunistic crimes and serial violent actions by taking advantage of opportunities they encounter in their activity spaces.

The goal of the project presented in [13] is to explore a methodology for reliably predicting location, time, and/or likelihood of future residential burglaries. First, a suitable data structure is designed to store spatial/temporal information as well as aggregated counts of crime and crime-related events categorized by the city's police department. Second, an ensemble of data mining classification algorithms is applied to perform residential burglary forecasting. In addition, it is also explored whether the crime rate will emerge or increase at certain locations (called 'heating up').

An algorithm to automatically detect patterns of crime (called Series Finder) is presented in [8]. Series Finder processes information similarly to the way crime analysts process information instinctively: the algorithm searches through the database looking for similarities among crime events and tries to identify the modus operandi (i.e., a frequent pattern) of a particular offender. The modus operandi is the set of habits that the offender follows, and is a type of motif used to characterize the pattern. As more crimes are added to the set, the modus operandi becomes more well-defined and the pattern should emerge as a frequent pattern of its criminal behavior.

In [5] a multivariate time series clustering technique based on dynamic time wrapping (DTW) and parametric Minkowski model is proposed to discover similar trends in crime data and subsequently use this information for future crime trends prediction. The algorithm has been tested on real-world datasets (provided by the Indian National Crime Records Bureau), performing a separated analysis for various types of crimes (i.e., murder, kidnapping, etc.)

The main issue investigated in [3] is whether it is possible to accurately forecast selected crimes on month ahead in small areas, such as police precincts. In a case study (Pittsburgh, PA) the forecast accuracy of univariate time series models (proposed in the paper) has been compared with naive methods (widely used by police). Moreover, Holt Exponential smoothing with monthly seasonality has been experimented using city-wide data and resulted as the most accurate forecast model for precinct-level crime series.

In [14] time series modeled by the autoregressive integrated moving average (ARIMA) methodology are used to make short-term forecasting of property crime in a city. The experimental evaluation consisted in training the model on a dataset collecting 50 weeks of property crime events, and through the model forecasting by the model the crime amount of 1 week ahead. The model's fitting and forecasting results are compared with the two exponential smoothing methods SES and HES. It is shown that the ARIMA model achieves higher fitting and forecasting accuracy than the other two methods.

In this work we propose an approach to forecast the number of crimes that will happen in rolling time horizons. Among the approaches proposed in literature, papers [14] and [3] share with our work similar issues and goals. However, with respect to those two papers, our work is new in two aspects. First, we show here the experimental evaluation performed on two-year-ahead crime forecasts, which (to the best of our knowledge) represents the longest time horizon considered so far in the literature. In fact, papers [14] and [3] report results on one-week-ahead and one-month-ahead predictions, respectively. Second, our approach shows higher predictive accuracy for long term crime forecasting. For example, in [3] it is reported a percentage error ranging between 13.7% and 33% (depending on the type of crime) on one-month-ahead forecasts, while [14] reports an accuracy of 90% in fitting the training set. The results reported in Section V show that our approach achieves a percentage error of 14.5% in one-month-ahead forecasts independently of the crime types, and an average error equal to 16% for one-year-ahead forecasts and 20% on two-year-ahead forecasts (that represents a time horizon 24 times longer than the longest one considered in the literature).

III. Background: auto-regressive models for time series forecasting

Multiple regression models have been defined with the goal of forecasting a variable of interest using a linear combination of predictors [15]. In particular, in an *auto-regression model*, the variable of interest is forecasted using a linear combination of its past values (the term auto-regression indicates that it is a

regression of the variable against itself), while a *moving average model* uses past forecast errors in a regression-like model. Sometimes, as a preliminary step to the regressive analysis, time series need a *differencing* transformation to stabilize the mean of a time series and so eliminating (or reducing) trend and seasonality. A combination of differencing, auto-regression and moving average methods is known as *AutoRegressive Integrated Moving Average* model (more frequently referred by its acronym *ARIMA*) [15], formally defined in the following.

Let us consider the time series $\{y_t : t = 1...n\}$, where y_t is the value of the time series at the timestamp t. Then, an $ARIMA(p, d, q)$ model is written in the form

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

where:

- $y_t^{(d)}$ is the d^{th}-differenced series of y_t, that is: $y_t^{(d)} = y_t^{(d-1)} - y_{t-1}^{(d-1)}, \ \ldots, y_{t-p}^{(d)} = y_{t-p}^{(d-1)} - y_{t-p-1}^{(d-1)}$;
- $\phi_1, \ldots, \phi_p$ are the regression coefficients of the auto-regressive part;
- $\theta_1, \ldots, \theta_q$ are the regression coefficient of the moving average part;
- $e_{t-1}, \ldots, e_{t-q}$ are lagged errors;
- e_t is white noise and takes into account the forecast error;
- c is a correcting factor.

The regression model above described is referred as $ARIMA(p, d, q)$, where the order of the model is stated by three parameters: p (order of the auto-regressive part), d (degree of first differencing involved) and q (order of the moving average part). The best parameter values are obtained by minimizing the BIC and AIC measures, as described in [15]. A useful notation commonly adopted when treating this kind of models is the 'backshift notation' [16], [17], [18], that is based on the B operator. The B (B^d) operator on y_t has the effect of shifting the data back one period (d periods). This is very useful when combining differences, as the operator can be treated using ordinary algebraic rules. By using the 'backshift' operator, the full model can be written as:

$$(1-\phi_1 B - \ldots - \phi_p B^p)(1-B)^d y_t = c + (1+\theta_1 B + \ldots + \theta_q B^q) e_t$$

whose details are out of the scope of this work and a formal demonstration can be found in [16], [17], [15].

Often, time series processes have a seasonal component that repeats with a given periodicity. In order to deal with seasonality, the classical ARIMA processes have been generalized and extended by the *seasonal ARIMA* models. 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 referred as $ARIMA(p, d, q)(P, D, Q)_m$, where m is a periodicity factor, and P, D and Q are the orders of the auto-regressive, differencing and the moving average part for the seasonal model, respectively [16], [17], [18], [15].

IV. Problem Definition and Goal

Let be $T =< t_1, t_2, \ldots, t_H >$ 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 be Y a time series of crimes, $Y =< y_1, y_2, \ldots, y_H >$, where each element y_i is the number of crimes occurred at the timestamp $t_i \in T$. Let us consider a future temporal horizon, $S =< t_k, t_{k+1}, \ldots >$, with $k > H$. Our goal is to find a regressive model for reliably predicting the number of crimes at a given timestamp $t_k \in S$. Formally, we want to extract a function $F_{ncrime} : \mathcal{S} \rightarrow \mathcal{R}$, that forecasts the number of crimes at a timestamp $t_k \in S$, in a given area of a city.

V. Crime Predictions: Analysis and Experimental Results

This section presents, as case study, the analysis of crimes occurring in an area of Chicago. The main goal consists in designing a predictive model to forecast the number of crimes that will happen in the future. In the following subsections we describe the main issues of our analysis: data description and gathering (SectionV-A), the regressive model training (SectionV-B), the residual analysis on the training set (SectionV-C) and the evaluation of the model on the test set (SectionV-D).

A. Data Description

Crimes data have been gathered from the *'Crimes - 2001 to present'* dataset, a real-life collection of instances describing criminal events occurred in the city of Chicago from 2001 to present (the repository is updated every week, so it is kept up-to-date minus the most recent seven days). Each crime is described by several attributes (i.e., type of crime, location, date, community area, etc.). This dataset has been retrieved as open data by the Plenario platform, a Web framework launched in late 2014, which gives public access to more than 100 datasets storing urban data [7]. Plenario was created (and currently managed) by the University of Chicago's Urban Center for Computation and Data (UrbanCCD)[1], as well as researchers from the Computation Institute[2] and Harris School of Public Policy[3].

As a pilot research study, in this work we focus our experiments on an area of Chicago, rather than performing an analysis on the whole city. The area chosen for our experiments, shown in Figure 1 and located in the North-East zone of the city, is growing in terms of population and business activities (so making it interesting for crime analysis). Its perimeter is about $18.144\ km$ and its area amounts to $10.869\ km^2$. Starting from the *'Crimes - 2001 to present'* dataset, we retrieved all crime events happened in the bounded area in 14 years (731 weeks), from January 2001 to December 2014. The total number of collected crimes is 174403, while the average number of crimes per week is 238 circa.

Figure 2 shows a time plot of the observed crime data, in which the number of crimes are plotted versus the time of observation. The plot immediately reveals some interesting

[1] www.urbanccd.org
[2] www.ci.uchicago.edu
[3] harris.uchicago.edu

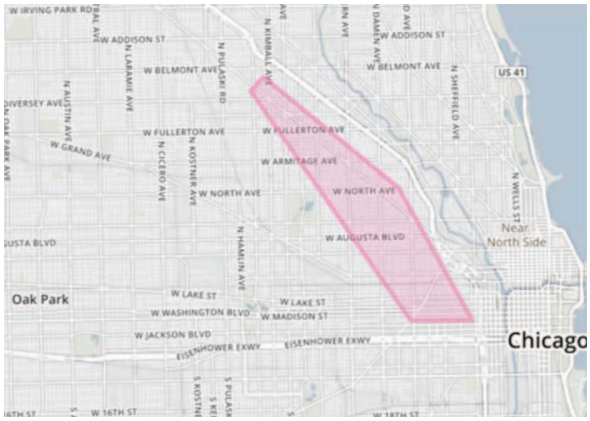

Fig. 1. Area of Chicago chosen to analyze crime trends and forecasts

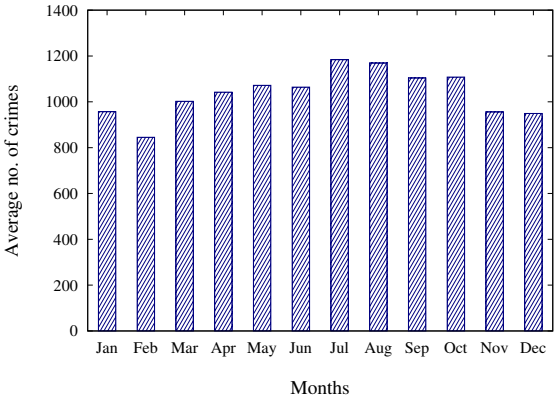

Fig. 3. Crime data set (2001-2014): distribution of number of crimes by month

features. First, it is evident that the number of crimes is decreasing with the years, showing a 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 level of the series decreases. From the plot, we can infer that the occurrences of crimes usually increase in the late Spring, achieve peaks during the Summer, decrease in Autumn and generally have dips in Winter. A better view of the seasonality hidden in the data can be appreciated in Figure 3, that shows the distribution of the average number of crimes by month. The histogram shows that the number of crimes happened in the city area under observation are highly skewed for different periods of the year. In particular, the number of criminal events achieves the highest values in July and August (with 1184 and 1169 crimes on average, respectively), while it strongly decreases in February (having a trough of 845 crimes, on average).

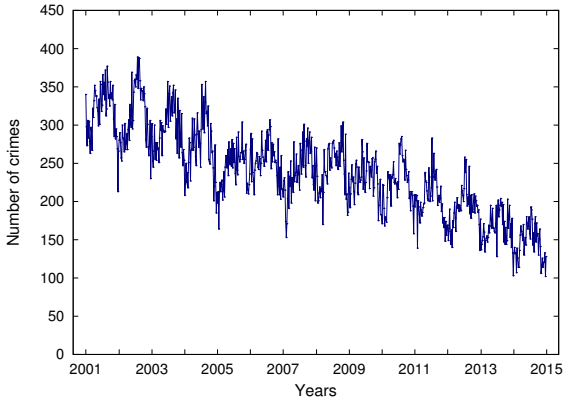

Fig. 2. Crime data set (2001-2014): number of crimes vs time

The *trend* and *seasonal pattern* can be highlighted by performing a decomposition task. In general, a seasonal time series y_t can be modeled as a linear combination of three components: a *trend-cycle* component (containing both trend and cycle), a *seasonal* component and a *remainder* component (containing anything else in the time series). To estimate those three components, generally two kinds of decomposition models are used: additive decomposition and multiplicative decomposition [15]. The first one is more appropriate when the magnitude of the seasonal fluctuations or the variation around the trend-cycle are stable with the level of the time series. The second one is more suitable when the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series (i.e., seasonal magnitude increases/decreases as the trend increases/decreases). Since the observed crime data belong to the second category, we performed a multiplicative decomposition to extract the components from the data [15]. The result of the time series decomposition is shown in Figure 4. That is composed of four panels: the top panel shows the observed data (i.e., original crime dataset), while the bottom three panels plot trend, seasonal and remainder components, respectively. If these three components are added together, we obtain the observed data shown in the top panel. It is worth noting that there is a strong decreasing trend, while the seasonal component is present with a similar pattern in all years.

As well known, to perform the regression task and its validation, we need to split the original dataset in two partitions: the training set and the test set. The first one is exploited to discover the relationships inside data while the second one is used for evaluating whether the discovered relationships hold. In our case, the whole crime time series has been split with respect to the number of years: the training set contains the time series of the first 12 years (2001-2012, 627 weeks), while the test set holds the time series of the last 2 years (2013-2014, 104 weeks). As described in the following subsections, we fit the regressive model using data from January 2001 to December 2012 and, by exploiting the trained model, we forecast the crime events from January 2013 to December 2014 to assess the quality of the predictions.

B. Training and Fitting the Regressive Model on the Training Set

As a first issue, we observe that the crime dataset under investigation (see Figure 2) is a no stationary time series. Conversely, a time series is defined to be stationary if its distribution does not depend on the time t. For such a reason, time series with trend or with seasonality (like the crime series we are analyzing) are not stationary, because trend and

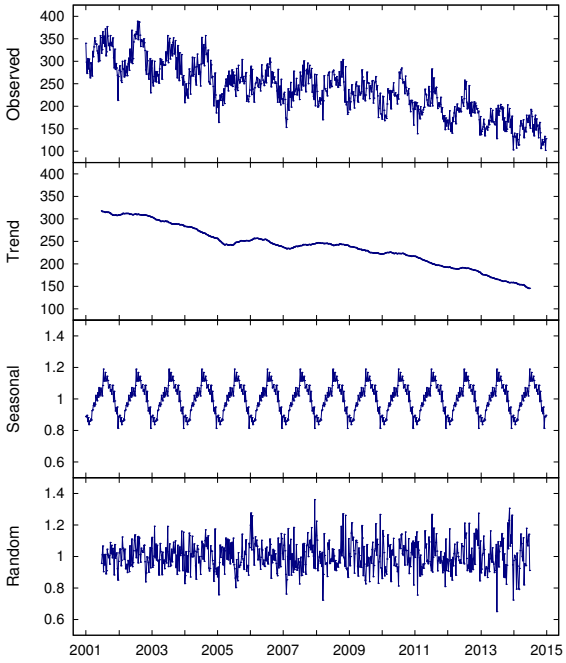

Fig. 4. Crime data set (2001-2014): observed, trend, seasonal and random components (from top to bottom)

seasonality can affect the value of time series at different times. This is an important issue to deal with, because ARIMA models are defined for stationary time series and they can not be applied for un-stationary data. Nevertheless, when we have a non-stationary time series, we can apply some transformation strategy to make it stationary. One of the most common techniques to make a time series stationary is differencing it, which can stabilize the mean of a time series, and so eliminating (or reducing) trend and seasonality. The differenced series is the change between consecutive observations in the original series, as described in Section III. Sometimes, it may be necessary to difference the data twice (or more times) until the differenced data do not appear stationary, in order to obtain a stationary series and thus to handle it by an ARIMA model. In our case, after performing a first differencing on the data, the differenced time series appeared to be stationary in its mean and variance, as well as its ACF tended to zero relatively quickly. For such reasons, we concluded that the first-order differenced time series is suitable for performing a regressive analysis by ARIMA. Consequently, we fixed $d = 1$ in the ARIMA model to be trained.

Once the differencing order has been chosen, the next step was to detect the optimal values of p and q (i.e., the number of auto-regressive and moving average regression coefficients, respectively) and the corresponding regression coefficients ($\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$). To do that, the values of p and q have been chosen by applying the Hyndman-Khandakar's algorithm, which is based on both the minimization of the AIC (Akaike's Information Criterion) and the maximization of the MLE (Maximum Likelihood Estimation) [15]. Briefly, the algorithm performs a stepwise search to traverse the model space and discover the optimal combination of p and q values, that can be explained in three main steps. First, several candidate models are obtained by a combination of p and q assuming values 0,1, and 2, and the model which minimizes the AIC is selected as the best current model. Then, variations on the best current model are considered by varying p and q by $\pm$ 1, and the AIC is computed for each one. If one candidate achieves a lower AIC than the best current model, then it becomes the new best current model. The procedure, iteratively, produces new variations of the best current model (varying p and q by $\pm$ 1) and computes their AICs. It terminates when there is no new candidate with a lower AIC than the best current model. In our case, we applied this algorithm on the crime data under investigation, and the best values resulted in $p = 1$ and $q = 1$. In a similar way, the best values for the seasonal part are chosen. In our case, the best values resulted in $P = 0$, $D = 1$ and $Q = 2$, with $m = 52$. Thus, the best auto-regressive model trained from the input data was $ARIMA(1,1,1)(0,1,2)_{52}$. Once the best ARIMA candidate has been identified, i.e., the triples (p, d, q) and (P, D, Q) values have been fixed, the estimation of the regression parameters of both seasonal (i.e., $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$) and non-seasonal part ($\Phi_1, \ldots, \Phi_p$ and $\Theta_1, \ldots, \Theta_q$) is obtained by the MLE computation, i.e., a methodology tending to maximize the probability of fitting the data that have been observed. At the end of the training step, we obtained the following parameter values: $\phi_1 = 0.1298$ and $\theta_1 = -0.9043$ for the non seasonal part, $\Theta_1 = -0.8974$ and $\Theta_2 = 0.0141$ for the seasonal part. The final autoregressive formula is formulated as:

$$(1-\phi_1 B)(1-B)(1-B^{52})y_t = (1+\theta_1 B)(1+\Theta_1 B^{52})(1+\Theta_2 B^{52})e_t \quad (1)$$

obtained with $MLE = 496.1$ and $AIC = -982.21$.

C. Residual analysis on the Training Set

Once the forecasting model was obtained (i.e., Equation 1), we executed a preliminary residual evaluation and diagnostics analysis on the training set. It is worth noting that a suitable forecasting method should yield residuals with the following properties: (i) they are uncorrelated (if there are correlations between residuals, then there is information left which should be used in computing forecasts), (ii) they have zero mean (if the residuals have a mean other than zero, then the forecasts are biased) and (iii) they are normally distributed and have constant variance (this last property is useful but not necessary). To this purpose, it is very useful to evaluate whether there are correlations between successive forecast errors, and whether forecast errors of the model are normally distributed with mean zero and constant variance.

To verify if the residuals satisfy the aforementioned requirements we exploit the autocorrelation function (ACF), that measures the linear relationship between lagged values of a time series and it is a useful tool to evaluate if the value observed at time t (current time) is influenced by the value assumed at time $t-k$ (k previous timestamps w.r.t. t). Figure 5 reports the correlogram of the forecast errors computed on the training set for the lags 1-30. It shows that only the lag 20 exceeds the significance bounds, so there is no significant correlation in the residual series (i.e., forecast errors). The lack of correlation suggests that the regressive model takes into account all the available information in the training set

data series. To understand whether the forecast errors are normally distributed with mean zero, we show in Figure 6 the distribution of the residuals (with overlaid normal curve with mean 0 and the same standard deviation as the distribution of forecast errors). The plot shows that the distribution of forecast errors is centered on zero and normally distributed. In conclusion, we can observe that successive fitting errors on the training data are not correlated and they are normally distributed with mean zero and constant variance. Obviously, the efficacy of the regressive model is determined not by its performance on the training data but by its ability to perform well on unseen data, i.e. the test set. For such a reason, in the next section we report on quality and accuracy of the regressive model evaluated on the test set, i.e. by considering data that were not used when fitting the model.

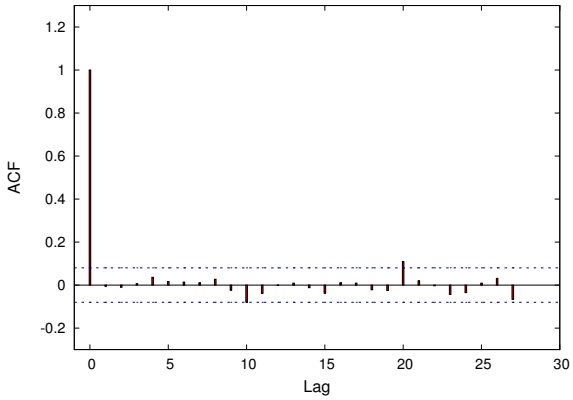

Fig. 5. ACF of the residuals computed on the training set versus the lags

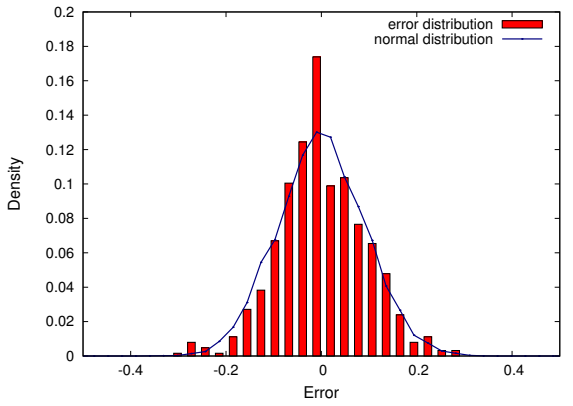

Fig. 6. Distribution of the residuals (with the overlaid normal curve) on the training set

D. Evaluating the Forecasting Model on the Test Set

To assess the effectiveness and accuracy of the regressive function modeled by Equation 1, we performed an evaluation analysis on the test set consisting of the last two years of data (i.e., 2013 and 2014). In particular, the model has been used to predict future values of the number of crimes that will happen in that area, week by week. The prediction of the type of crimes, as well as the locations they will happen, is out of the scope of this work and it will be studied in a further research activity.

Let y_i denote the i^{th} observation and $\widehat{y_i}$ denote the forecast of y_i according to the formula in Equation 1 for $t = t_i$. Figure 7 shows four curves versus the time ranging from 2001 to 2014. For the training set period, from 2001 to 2012, observed data and fitted data are plotted in black and red, respectively. For the test set period, i.e. from 2013 to 2014, observed and forecasted data are traced in blue and green, respectively. It is interesting to highlight that the regressive curve fits well the training data series, with an overfitting only in the first year (where the black line is hidden by the red line). By looking at the test set, we can notice that forecasted data adhere very well to the observed data for that period. First, it is evident that the decreasing trend forecasted by regressive model is very similar to that one occurring in the observed data. Second, it is impressive how the seasonal patterns within each year (and their corresponding maximum and minimum peaks) are well modeled by the regressive model. For example, the peak of crimes really occurred in the 27^{th} week of 2014 (193 crimes) were forecasted, by the regressive model, to happen in the 28^{th} (211 crimes); similarly, the trough of crimes, which occurred during the 51^{th} week of 2014 (102 crimes) was forecasted to happen in the 50^{th} of the 2014 (148 crimes). For a more detailed view, Figure 8 shows observed and forecasted data of the test set. We can notice that predicted values in general are a bit higher than observed data, by showing an over-forecasting with respect to the real number of crimes.

Now, let us give a quantitative evaluation about the accuracy of the regressive model. To do that, we considered two measures: the forecast error $e_i = y_i - \widehat{y_i}$ and the percentage forecast error $p_i = \frac{e_i}{y_i} \cdot 100$. Based on such measures, we computed several indices (commonly used in literature) to evaluate the forecasting accuracy:

- *Mean Absolute Error*: $MAE = mean(|e_i|)$, i.e., a scale-dependent index measuring the average forecasting absolute error;
- *Mean Absolute Percentage Error*: $MAPE = mean(|p_i|)$, i.e., a scale-independent index computing the average forecasting percentage error;
- *Mean Error*: $ME = mean(e_i)$, i.e., a scale-dependent index measuring the average forecasting relative error;
- *Root Mean Squared Error*: $RMSE = \sqrt{mean(e_i^2)}$, i.e., a general purpose error metric for numerical predictions that amplifies and severely punishes large errors.

Table I reports the values of the four indices described above, for the test set (years 2013 and 2014). More in detail, we computed the error measured by considering several time horizons for the test set: the two semesters of 2013 and 2014, the whole years 2013 and 2014, and finally the complete period 2013-2014 (corresponding to the total test set). Looking at the values in the table, we can do three main observations. First, forecasting errors increase when the prediction horizon is longer and longer. For example, the MAE monotonously increases from 23.88 (for the first semester of 2013) to 37.38 (for the second semester of 2014), and similarly all other indices. This is a reasonable result, because the predictions

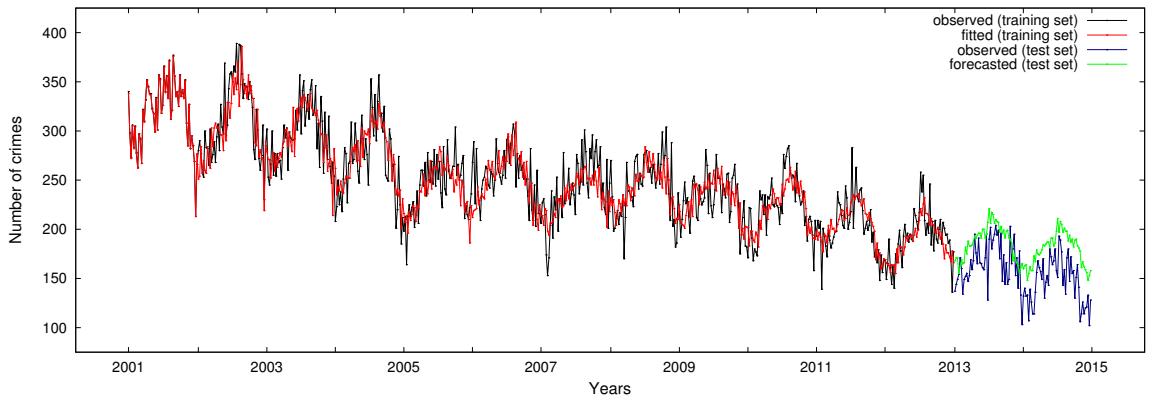

Fig. 7. Number of crimes observed and fitted (black and red lines) on the training set, and number of crimes observed and forecasted (blue and green lines) on the test set

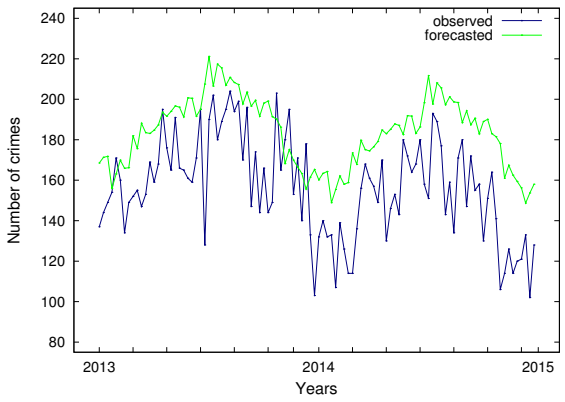

Fig. 8. Number of crimes observed and forecasted (blue and green lines) on the test set

depend on the previous historical trends: more distant is the forecasting time from the last historical data, lower is the forecasting accuracy. Second, percentage errors (MAPE column) show that the adopted regressive model (Equation 1) forecasts the number of crimes with an average error of 16% for the first year , that appears to be a very interesting result. If the prediction horizon is extended to two years, the percentage error increases to 20%. Third, we observe that the mean error (ME column) assumes negative values in all time windows, confirming that predicted values are on average higher than observed data. This means that the regressive model tends to over-forecast the number of crimes with respect to those will happen in reality. To better understand this issue, Figure 9 shows the histogram of the forecast errors of the whole test set, with an overlaid normal curve with mean 0 and the same standard deviation as the distribution of errors. The plot shows that the distribution of forecast errors is shifted towards negative values compared to a normal curve (it should be centered on 0, in the ideal case). As reported in Table I, the mean error for the whole test set period (2014 and 2015) is -27.05: this is absolutely consistent with the residual histogram in Figure 9.

Figure 10 shows the forecast error e_i versus the number of weeks in the test set, with $i = 1, \ldots, 104$. The highest error

TABLE I. FORECAST ERROR MEASURES VS SEVERAL TIME WINDOWS, FOR YEARS 2013 AND 2014

Time Window	MAE	RMSE	MAPE	ME
Jan-Jun 2013	23.88	28.86	0.16	-23.17
Jul-Dec 2013	25.07	29.62	0.16	-19.58
Jan-Jun 2014	28.09	31.58	0.21	-28,09
Jul-Dec 2014	37.38	41.17	0.28	-37.38
Year 2013	24.47	29.25	0.16	-21.37
Year 2014	32.73	36.69	0.24	-32.73
Years 2013-2014	28.61	33.18	0.20	-27.05

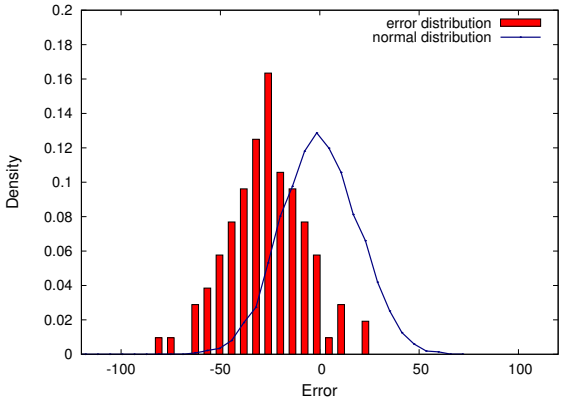

Fig. 9. Distribution of the residuals (with the overlaid normal curve) on the test set

value is 22, occurring in the 50^{th} of 2013, when the number of observed crimes was 178 and the forecasting was 156. The average error amounts to -27.05, confirming that forecasted values are on average higher than the observed values (as previously reported in the comments of Figure 8).

Figure 11 shows the absolute percentage error $|p_i|$, versus the number of weeks in the test set. We observe that $|p_i|$ increases with the number of weeks. In fact, we can observe that in the first 52 weeks (first year of prediction, i.e., 2013) the average percentage error is lower than in the second 52 weeks (second year of prediction, i.e., 2014). It is worth noting that in the final time window of the test set, i.e., the last 26 weeks, $|p_i|$ tends to strongly increase with respect to the time horizon.

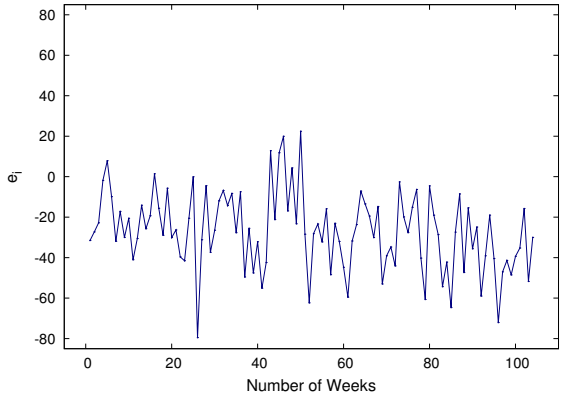

Fig. 10. Forecast error e_i versus the number of weeks in the test set

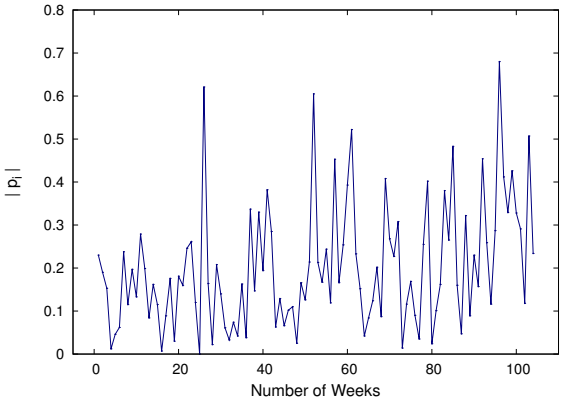

Fig. 11. Absolute percentage error $|p_i|$ versus the number of weeks in the test set

VI. Conclusion

New technologies are enabling police departments to access growing volumes of data related to crime, that can be analyzed to understand patterns and trends. Such knowledge is useful to anticipate criminal activity and predict/prevent crimes. This paper presented an approach, based on auto-regressive models, for reliably forecasting crime trends in an urban area. Experimental evaluation, performed on crime data of an area of Chicago, showed that the proposed methodology can forecasts the number of crimes with an accuracy of 84 percent on one-year-ahead and 80 percent on two-year-ahead forecasts. At the best of our knowledge, such exceeds those of other approaches proposed in the crime forecasting literature.

In future work, several research issues may be investigated. First, we may perform an extended experimental evaluation on other urban territories, to assess the results obtained in the case study reported here. Second, we may correlate the trend of crimes and other events of the city, to discover any correlations among those. Third, in addition to the number of crimes, it may be interesting to investigate some methodologies that predict the type of crimes and the area they will likely happen.

Acknowledgment

This research was partially supported by the MIUR projects DICET-INMOTO (PON04a2_D) and DOMUS (PON0050_2). We are grateful to Albino Altomare, of ICAR-CNR, for his assistance and help given during the experimental evaluation. We also thank Brett Goldstein and Maggie King, of the University of Chicago's Harris School of Public Policy and Computation Institute, for their advice.

References

[1] United Nations Settlements Programme, the state of the world's cities 2004/2005: Globalization and urban culture. Earthscan, 2004.

[2] D. E. Brown and S. Hagen, "Data association methods with applications to law enforcement," *Decision Support Systems*, vol. 34, no. 4, pp. 369–378, 2003.

[3] W. Gorr, A. Olligschlaeger, and Y. Thompson, "Short-term forecasting of crime," *International Journal of Forecasting*, vol. 19, no. 4, pp. 579 – 594, 2003.

[4] M. Tayebi, M. Ester, U. Glasser, and 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] B. Chandra, M. Gupta, and 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, pp. 892–896.

[6] E. Cesario, C. Comito, and D. Talia, "Towards a cloud-based framework for urban computing, the trajectory analysis case," in *Third International Conference on Cloud and Green Computing (CGC'2013),*, 2013, pp. 16–23.

[7] C. Catlett, T. Malik, B. Goldstein, J. Giuffrida, Y. Shao, A. Panella, D. Eder, E. van Zanten, R. Mitchum, S. Thaler, and I. T. Foster, "Plenario: An open data discovery and exploration platform for urban science," *IEEE Data Eng. Bull.*, vol. 37, no. 4, 2014.

[8] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Learning to detect patterns of crime," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, 2013, pp. 515–530.

[9] J. S. d. Bruin, T. K. Cocx, W. A. Kosters, J. F. J. Laros, and J. N. Kok, "Data mining approaches to criminal career analysis," in *Proceedings of the Sixth International Conference on Data Mining*, ser. ICDM '06, 2006, pp. 171–177.

[10] G. Wang, H. Chen, and H. Atabakhsh, "Automatically detecting deceptive criminal identities," *Commun. ACM*, vol. 47, no. 3, pp. 70–76, 2004.

[11] S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, 2006, pp. 41–44.

[12] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[13] C.-H. Yu, M. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 779–786.

[14] P. Chen, H. Yuan, and 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] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts.com, 2014.

[16] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, 3rd ed., ser. Springer Texts in Statistics. New York: Springer, 2011.

[17] P. S. P. Cowpertwait and A. V. Metcalfe, *Introductory Time Series with R*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[18] J. Cryer and K. Chan, *Time Series Analysis: With Applications in R*, ser. Springer Texts in Statistics. Springer New York, 2008. [Online]. Available: https://books.google.it/books?id=bHke2k-QYP4C