A spectral analysis of crimes in San Francisco

A spectral analysis of crimes in San Francisco

Luca Venturini
luca.venturini@polito.it

Elena Baralis
elena.baralis@polito.it

Politecnico di Torino
Dipartimento di Automatica e Informatica
Corso Duca degli Abruzzi 24, Torino, Italy

ABSTRACT

In this work, we attempt an exploratory analysis of spatio-temporal patterns of crime in San Francisco. We apply spectral analysis to the temporal evolution of all categories of crime, finding that many have a weekly or monthly periodicity, along with other components. We show that spatial distribution has weekly patterns as well. These results can improve our understanding of the dynamics of crime and be exploited to design predictive models for policing.

CCS Concepts

•Information systems → Spatial-temporal systems; *Data mining;* **•Applied computing → Sociology;**

Keywords

Urban analysis; time-series; seasonality; crime; Lomb-Scargle periodogram

1. INTRODUCTION

In the last decades, smart cities and administrations have released large amounts of data as Open Data. First pulled by a part of the citizenry advocating for transparency, Open Data have been lately pushed and supported by the same administrations, willing to sustain new studies and foster innovative applications of the data, as integrated visualizations or predictive models. In the years, wide, comprehensive and structured datasets of heterogeneous categories have been growing to dimensions that now allow significative insights and powerful applications.

Urban data, being strictly linked with human activities, usually hide repetitive patterns, which occur over time and space. The exploration of these patterns can follow the intuition of the scientist, inspired by a shared knowledge of the topic or by the literature, or be the fruit of a systematic approach of data mining. Highlighting reoccurring patterns gives valuable insights into the data and thus increases the knowledge on the subject, laying also the basis for predictive or explanatory models.

In this work, we aim at finding spatio-temporal patterns in a large dataset of urban crimes, through spectral analysis. Our research question is twofold: firstly, we want to find out whether seasonal patterns exist in the data and can be found, and secondly, whether such patterns are global or if they vary by the category of crime. The answer to these questions can improve the current understanding of these phenomena and drive our future research on a predictive model for policing. To this aim, we propose a methodology for mining seasonal components in a time series that we borrow from signal processing theory. We couple this technique with heatmap analysis to lay down an exploratory process able to highlight patterns in time and space.

2. TIME-SERIES ANALYSIS

2.1 Description of the dataset and motivations

The dataset used in our analysis describes all the crime events in the city of San Francisco from January 2003 to August 2016, as published by the San Francisco Open Data portal [1]. The dataset counts 1952810 records, split in 39 major categories of incidents, e.g. Assault, Vehicle Theft, Drug/Narcotic, and each record can be assigned to one of these or to Other Offenses. Some incidents can consist of several records, as in the case of multiple, contemporary arrests or of multiple charges on the same person. Each record reports also a short description of the incident and its resolution, namely if it led to an arrest. The metadata available are the timestamp and the location, with the address, the district and the spatial coordinates of the event.

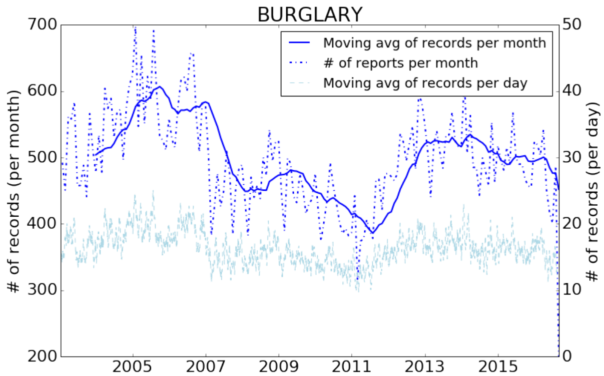
The richness of this dataset spans multiple dimensions and allows a plethora of interesting insights and analyses. Here we focus primarily on the temporal aspect, and we show in Figure 1 a sample of the width of the data available, already narrowed down to a single category, i.e. Burglary. Figure 1(a) shows a glimpse on the more of 13 years of incidents recorded, with different scales of aggregation. The plot of the number of records per month, with its respective trend, computed as a moving average with a 12-month window, already shows how such events can hugely vary within months, seasons or years. The hundreds of records per month and the tens per day allow thorough, statistically significant analyses, on different scales and resolutions.

In Figure 1(b) we show a detail over the daily distribution, to highlight how the distribution of the events greatly varies not only from one month to the other, but also within conse-

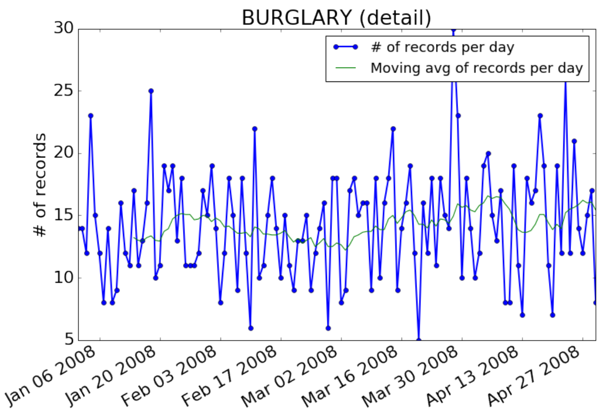
UrbanGIS 16, October 31-November 03 2016, Burlingame, CA, USA

ACM ISBN 978-1-4503-4583-5/16/10…$15.00

DOI: http://dx.doi.org/10.1145/3007540.3007544



(a) Temporal evolution of burglary from 2003 to 2016



(b) Daily evolution of burglary in 2008 (detail)

Figure 1: Evolution of burglaries in San Francisco

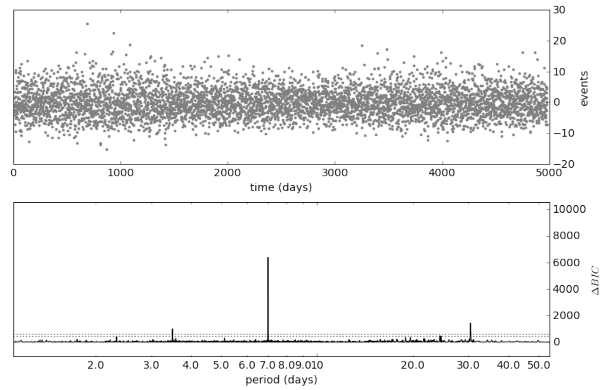


Figure 2: Periodogram of burglaries with detrending

quent days and weeks. Thus, a model aiming at predicting or explaining the evolution of the incidents in the city would need to cope with a great complexity, result of several hidden seasonal components, each one potentially peculiar of only some categories of crime, added to the natural variation of the datum.

2.2 Spectral analysis

In order to highlight the seasonal components of the dataset we resort to the Lomb-Scargle periodogram [9, 11]. Periodograms, like the Lomb-Scargle, show the most likely periods in a time series and are thus a kind of spectral analysis. The Lomb-Scargle periodogram is usually preferred to the more known Fourier transform when dealing with missing samples or uneven sampling steps, and it is therefore of much wider application.

This technique is also known as least-squares spectral analysis, as it is a least-square fit of sinusoidal functions over the data. The peaks found in the resulting periodogram are thus the periods which best fit the dataset, with the y-axis showing the gap in the Bayesian Information Criterion, a measure of how likely one period fits better than other values.

The Lomb-Scargle periodograms in our experiments were computed using the Python package AstroML [13].

3. RESULTS

Before applying the Lomb-Scargle analysis, we isolate the seasonal components of the series. We do this by subtracting the trend, computed with a simple 14-day moving average of the data itself. Figure 1(a) shows this moving average for burglaries. The result of the decomposition is shown in the upper part of Figure 2, a scatter plot of the detrended signal, whose central value is zero, as wanted. Below this signal, we see its periodogram, computed on a set of candidate values ranging from 1.1 to 50 days, to find short and medium-term periods. The periodogram evidences three significant periods for burglaries: one at seven days, one at 3.5 days, a sibling of the main peak, and one at a value between 30 and 31 days, which corresponds most probably to the average length of a solar month. The prominent period for burglaries is by far the one at 7 days.

We applied the Lomb-Scargle analysis to all the 39 categories of records, to find temporal patterns in the dataset and to state whether a unique predictive model can fit any category of crime, and thus all records, or if instead each category needs a different model. Several categories do not have more than some hundred samples, though, and therefore do not produce significant periodograms. In Figure 3 we show the periodograms for the 14 most-frequent categories of crime. Among them, we see very specific categories like Vehicle Theft, and broader categories as Other Offenses, Non Criminal and Secondary Codes. The latter did not show significant periods, most probably because of the variety of events that belong to it. On the other hand, the other very assorted category, that is Other Offenses, is depicted by a very neat peak at 7 days, with no sign of other possible periodicities. Also almost all the other categories show this clear weekly period. The second most represented period is the monthly one, which is discernible in at least half of these 14 categories. As seen with burglaries, this period is always slightly more than 30 days. This result excludes the hypothesis of a 4-week pattern, and suggests the idea that these crimes tend to happen at the same day of the month. Robberies, interestingly, as opposite to all other kind of crimes, seem to have such monthly patterns way more likely than the weekly ones. Also interesting is the fact that we never have periods of 14 days, and only 3 periodograms show peaks at 15 days. These results encourage the investigation of seasonal modelings of the temporal evolution of crimes and suggest to focus the attentions on weekly and monthly cycles primarily.

We now consider the spatial distribution of the events for one of the major categories, Vehicle Theft. The periodogram drives us towards an inspection of weekly patterns, but still does not say if the patterns are homogeneous in space, or peculiar to some parts of the city. Figure 4 shows a heatmap of all the vehicle thefts recorded in 2015, aggregated by day of the week. Thursdays, not shown here for space constraints [1], are most similar to Wednesdays. The two central weekdays are thus where the hotspots of events are most extended, with a slight relaxation in Mondays and Saturdays. Like in the time series, we can glimpse here a trend, in the form of some core spots common to all days, and some components which are peculiar to each day. Focusing on the hottest zones (in red in the plot) of Wednesdays and Sundays, for example, we see in both two neat clusters, which are though different in shape, position and extension. These trends might suggest that analyses similar to what we have tried on time, are possible also on spatial dimensions.

4. RELATED WORK

Seasonality in criminal incidents has been widely studied. Works like [10] and, more recently, [2, 4, 5, 8] have already attempted a reasoning over the reproduction of crimes on regular cycles. [5] considers the seasonal component of climatic seasons, aiming primarily at finding differences correlated with weather conditions and climate. Similarly, [4, 8] deal with the link between some categories of crime, like assaults or property crimes, and times of the year, searching for peaks and correlations with external factors. Our approach instead is agnostic towards any initial hypothesis of seasonality or correlations and comprehensive of all categories of crimes. Closer to this effort is [2], which studies the seasonality of different kind of crimes, but always with a focus on month periods. In our work we instead focused on short and medium-term cycles of few days or weeks, that could be helpful for short-term predictions. The availability of full timestamps for crimes is a must for a finer resolution of the analysis, e.g. at the day scale; works like [2, 8], for example, rely on the Vancouver open dataset, which exposes only the year and month of the happenings.

The usage of Lomb-Scarge periodograms in a scope of social interest is, to our knowledge, a novelty. Since its introduction in 1976 [9], it has been applied in its birth domain of astronomical studies [11], seismology [3], and, more recently, biology [12, 7], even though spectral analysis is a wide-spread technique in many scientific domains.

5. FUTURE DIRECTIONS

In this work, we have analyzed the open dataset of San Francisco crimes and its temporal evolution, seeking for seasonal patterns that could be helpful for predictions. The spectral analysis has brought to evidence a number of interesting insights that were not immediately clear from a simple look at the curves, hidden by the complexity of the data. For example, we discovered the tendency of some categories of crime to repeat on a monthly basis, like robberies, and how almost all show a weekly period, like vehicle thefts.

Further investigations would need to drill down the analyses to finer granularities of time and space, e.g. hours and districts, to verify the stationarity of the findings at smaller scale. The need for a consistent number of observations may limit the analyses, though, to the districts with higher frequencies of crimes, where they are anyway more useful.

We plan to exploit our findings in the design of an algorithm aimed at predicting near-future crime events, similarly to [6]. The discovered temporal seasonalities support the design of models based on weekly and monthly patterns and features chosen to highlight such patterns, e.g. the number of same-kind events in the neighborhood in the same day of the week or in the same day of the month in the last year. In the model we will take into account spatial patterns and neighborhood-level deviations from the global trends, and the integration of external data.

[1] The complete plots of this paper, together with the code and data to reproduce them, are available at https://github.com/lucaventurini/timecrime

6. REFERENCES

[1] San francisco open data portal. https://data.sfgov.org/. Accessed: 2016-09-01.

[2] M. A. Andresen and N. Malleson. Crime seasonality and its variations across space. *Applied Geography*, 43:25–35, 2013.

[3] S. Baisch and G. H. Bokelmann. Spectral analysis with incomplete time series: an example from seismology. *Computers & Geosciences*, 25(7):739–750, 1999.

[4] G. D. Breetzke and E. G. Cohn. Seasonal Assault and Neighborhood Deprivation in South Africa: Some Preliminary Findings. *Environment and Behavior*, 44(5):641–667, 9 2012.

[5] V. Ceccato. Homicide in São Paulo, Brazil: Assessing spatial-temporal and weather variations. *Journal of Environmental Psychology*, 25(3):307–321, 2005.

[6] A. Chohlas-Wood, A. Merali, W. Reed, and T. Damoulas. Mining 911 Calls in New York City: Temporal Patterns, Detection and Forecasting. 2015.

[7] E. F. Glynn, J. Chen, and A. R. Mushegian. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics*, 22(3):310–316, 2 2006.

[8] S. J. Linning. Crime seasonality and the micro-spatial patterns of property crime in Vancouver, BC and Ottawa, ON. *Journal of Criminal Justice*, 43:544–555, 2015.

[9] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462, 2 1976.

[10] L. R. McPheters and W. B. Stronge. Testing for seasonality in reported crime data. *Journal of Criminal Justice*, 1(2):125–134, 22 1973.

[11] J. D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835, 12 1982.

[12] H. P. A. Van Dongen, E. Olofsen, J. H. VanHartevelt, and E. W. Kruyt. Searching for Biological Rhythms: Peak Detection in the Periodogram of Unequally Spaced Data. *Journal of Biological Rhythms*, 14(6):617–620, 12 1999.

[13] J. VanderPlas, A. J. Connolly, Z. Ivezic, and A. Gray. Introduction to astroML: Machine learning for astrophysics. In *2012 Conference on Intelligent Data Understanding*, pages 47–54. IEEE, 10 2012.

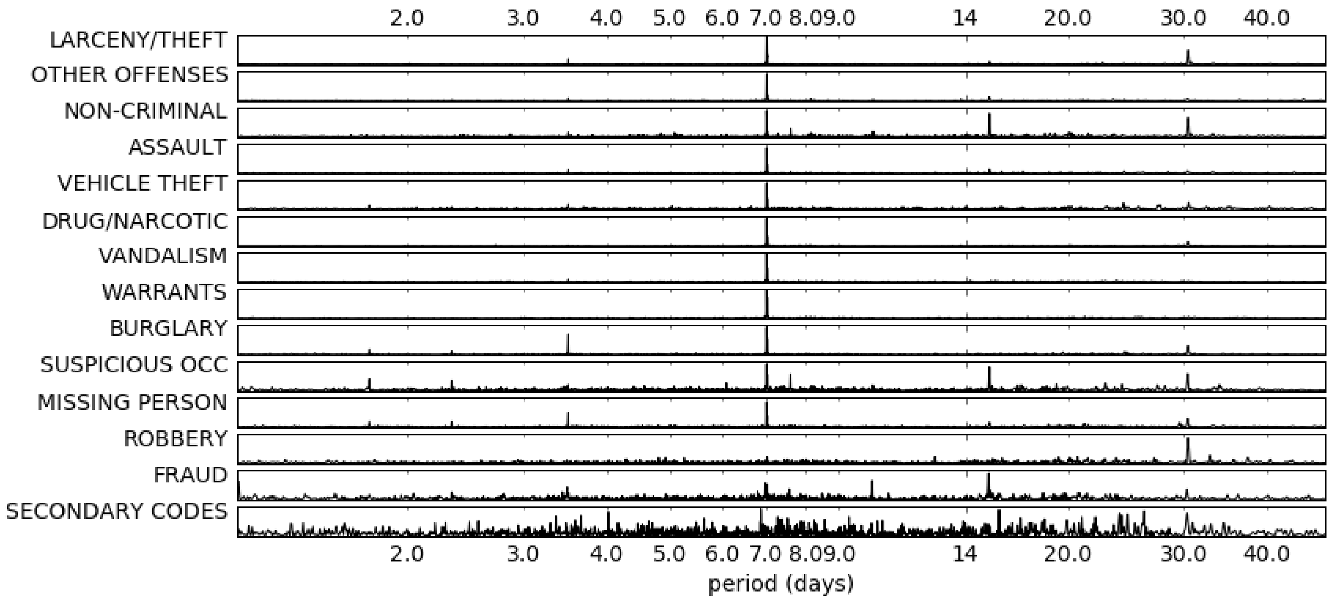


Figure 3: Periodogram of most frequent categories of crime

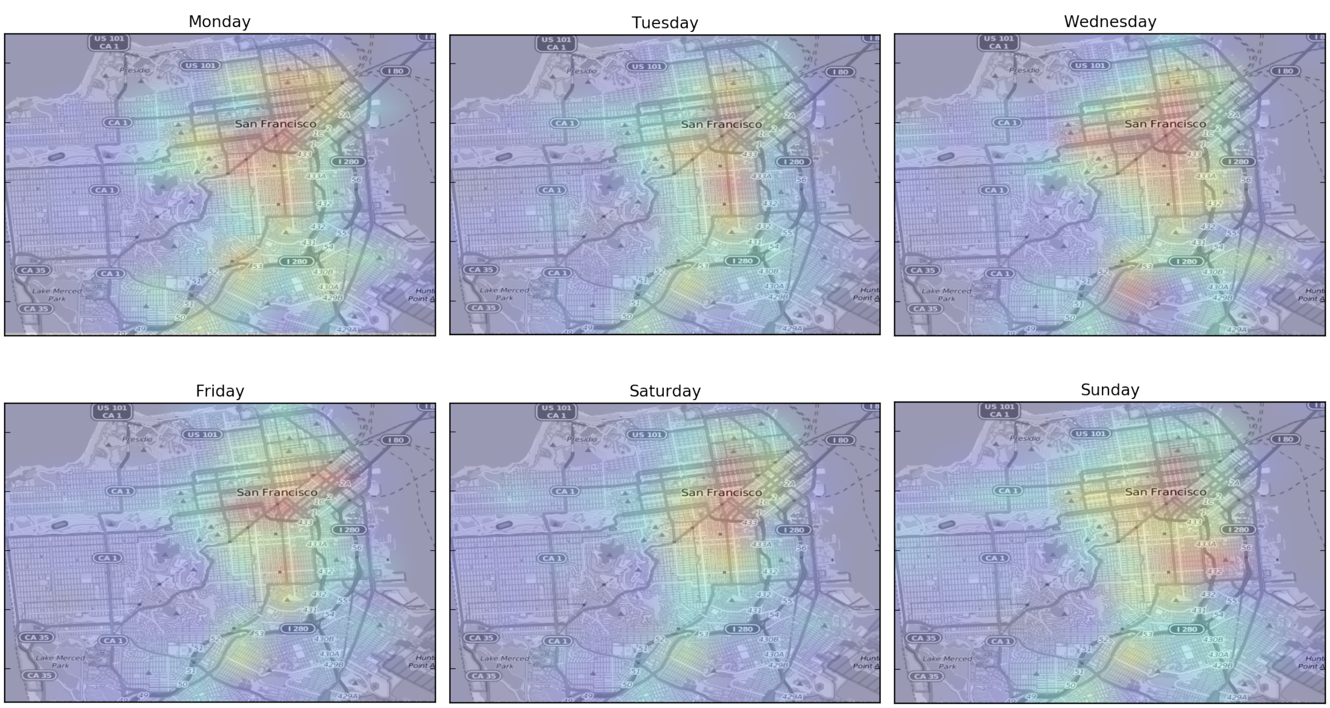


Figure 4: Heatmap of vehicle thefts in different days of the week, in 2015