Journal of Big Data

CASE STUDY

Open Access

Exploring crime patterns in Mexico City

C. A. Piña-García[1*] and Leticia Ramírez-Ramírez[2]

*Correspondence:
cpina@uv.mx
[1] Laboratorio para el Análisis de Información Generada a través de las Redes Sociales en Internet (LARSI), Centro de Estudios de Opinión y Análisis, Universidad Veracruzana, Xalapa, Mexico
Full list of author information is available at the end of the article

Abstract

Introduction: Life in the city generates data on human behavior in many different ways. Measuring human behavior in terms of criminal offenses plays a critical role on identifying the most common crime type in each urban area. A primary concern for the population of the largest cities in the world is to avoid specific city zones that could show a significant risk. This case of study systematically explores different sources of data including social media related to crime reports. In addition, this research seeks to examine and predict the most frequent crimes in Mexico City.

Case description: This case study was conducted in the form of an exploratory research. One of the parties involved is the Mexico City Police Department which released the approximate locations and categories of all crime reports from January 2013 to September 2016. We analyze the impact of crime in Mexico City based on 13 official crime categories: Robbery passerby, Theft of motor vehicle, Robbery of business property, Card fraud, Homicide, Domestic burglary, Robbery on public transportation, Rape, Firearm injuries, Robbery in subway, Robbery on taxi, Robbery to carrier, and Robbery to deliver person. We compare and analyze how people report a crime through the traditional system and using social media.

Discussion and evaluation: This research uses a quantitative case study approach to investigate, how our predictive model is able to forecast the total number of reported crimes in the following week based on its previous weekly aggregated observations and Google Trends series. Similarly, this case study seeks to determine whether Twitter performs correctly as a "social crime sensor" in terms of detecting certain areas and boroughs that are more likely to show criminal behavior.

Conclusions: In this study we used a linear predictive model in order to evaluate the performance of Google Trends in predicting crime rates based on weekly analysis. In addition, Twitter showed a suitable performance to discover the spatial distribution of crime frequency in Mexico City. Finally, this study provides an important opportunity to develop and encourage tailored strategies to tackle crime.

Keywords: Big Data, Crime patterns, Mexico City, Predictive model, Twitter, Google Trends

Introduction

With its more than 20 million citizens Mexico City is considered one of the largest cities in the world in terms of population and density. This city is divided in 16 boroughs which enjoy a certain degree of political autonomy from the city administration. These boroughs are listed as follows: Álvaro Obregón, Azcapotzalco, Benito Juárez, Coyoacán, Cuajimalpa, Cuauhtémoc, Gustavo A. Madero, Iztacalco, Iztapalapa, Magdalena

Contreras, Miguel Hidalgo, Milpa Alta, Tláhuac, Tlalpan, Venustiano Carranza, and Xochimilco. Similarly, these boroughs are further segmented into hundreds of neighborhoods. Having in consideration the size of this metropolitan area, it can be said that crime in Mexico City is very hard to measure. Police reports generally understate crime substantially and it can be extremely misleading. Therefore, recorded crime statistics need to be treated with great caution [1, 2]. However, police reports are usually the only available approach to gauge local crime.

It is important to note that this situation is derived from a high rate of criminal acts such as robberies, card frauds, homicides and various street crimes. Thus, Mexico City is continuously experiencing specific conditions that collectively degrade the security environment in certain areas. According to the Procuraduria General de Justicia de la CDMX [3] and the Secretaría de Seguridad Pública de la CDMX which is charged with maintaining public order and safety in Mexico City [4], the following city boroughs routinely show the highest number of crimes reported since 2014: Iztapalapa, Cuauhtémoc, Gustavo A. Madero, Benito Juárez, Coyoacan, and Tlalpan.

The Mexico City Police Department recorded 132,692 offenses from January 2013 to September 2016; these reports suggest that the police is dealing with a growing volume of crime. This research is intended to provide information on short-term trends alongside additional online data on the characteristics of victims and nature of crime.

Questions have been raised about the actual or estimated reports related to crime. The actual crime rate is thought to be much higher due to the fact that most people are reluctant to report a crime (only one out of 100 reported crimes actually goes to sentencing). It should be noted that nine out of ten crimes remain unreported in Mexico City, this is due to a lack of trust from citizens to the authorities. Thus, public-opinion polls have showed that the security problem remains the central concern of most citizens [5, 6]. In this regard, the main challenge faced by many researchers is the difficulty to get reliable crime data.

There is an urgent need to address the safety problems caused by crime. However, the lack of trust in authorities and insufficiency of data has been a controversial issue for many years. To date, apart from official crime reports there has been no reliable evidence that indicates that crime is decreasing. Therefore, despite the effort of the government of Mexico City in terms of crime prevention strategies, there still remains a paucity of information that permits to identify a solid measure of all crime.

Few studies have investigated crime concentration in any systematic way [7–10]. So far, however, there has been little discussion about the specific situation in Mexico City. Previously published studies on the subject has been mostly restricted to activities concerning illegal drugs and gang-related violence [11, 12]. In this sense, to the best of our knowledge, there has been hardly anything crime density analysis in Mexico City.

In recent years, there has been an increasing interest in online social media monitoring and Big Data analytics. Social media platforms are being used to gather information and in some cases, to examine crime prediction [13–15]. However, much of the research up to now has only focused on cities from USA such as Chicago, Arizona, San Francisco and New York [16–19].

Moreover, Google Trends has already drawn attention in many real-world phenomena e.g., epidemiology, economy, political science and crime prediction. In [20] Google

Trends Data was used to predict methamphetamine-related crime in Switzerland, Germany, and Austria. Similarly, a number of studies have begun to examine the combined use of Twitter and Google Trends for decisions making [21]. Nevertheless, despite the importance of Mexico City in terms of crime and antisocial behavior, this high crime rated city has not been investigated properly from the social media and Big Data perspective.

In this regard, there has been no relevant evidence that provides a good measure of short-term trends for a selected range of crimes experienced by individuals, including those not reported to the police. Official reports are still the only source to assess the impact of criminal behavior among local residents. The importance and originality of this case of study is that it explores and compares crime reports using a Big Data and social media approach. This approach is based on statistical methods and big data collection techniques to enhance our crime perception in Mexico City through data analysis and predicting models.

The rest of the paper is structured as follows: "Case description" section will provide a suitable overview of official crime data in Mexico City. The "Data and methods" section will describe how was the data collected for this manuscript. Preliminary results will be given in a systematic and detailed way in "Results and discussion" section. Finally, a summary explaining the significance of the findings will be highlighted in "Conclusions" section.

Case description

Crime by its nature is inherently difficult to measure it, and there is none reliable methodology able to capture a complete picture of this social problem [22]. However, official reports provide the best possible way to explore and study this type of social issue. The Mexico City Police Department released the approximate locations of 13 crime categories: (1) Robbery passerby, (2) Theft of motor vehicle, (3) Robbery of business property, (4) Card fraud, (5) Homicide, (6) Domestic burglary, (7) Robbery on public transportation, (8) Rape, (9) Firearm injuries, (10) Robbery in subway, (11) Robbery on taxi, (12) Robbery to carrier, and (13) Robbery to deliver person. This dataset can be freely downloaded from [23].

Preliminary analysis of this data shows the most common crime per borough (local crime patterns). This information can be gridded to produce choropleth maps showing the approximate locations in terms of latitude and longitude where a crime was reported as is shown in Figs. 1, 2. These maps provide the number of crimes, a single-hue progression fade was used in this map where the darkest hue represents the greatest number in the data set and the lightest shade represents the least number.

Similarly, Fig. 3 presents a heatmap with the relative intensity of crime occurrence i.e., crimes with a highest frequency in their value relative to the others will be given a “hot” color (high level), while those crimes that are lower in their value relative to the others will be given a “cold” color (low level). From figures described previously, it can be seen that the inner part of the city shows a high level of crime, while further out in the suburbs the level of crime slightly decreases. Note that this is a map that does not distinguish between high-crime and low-crime areas. However, it is important to

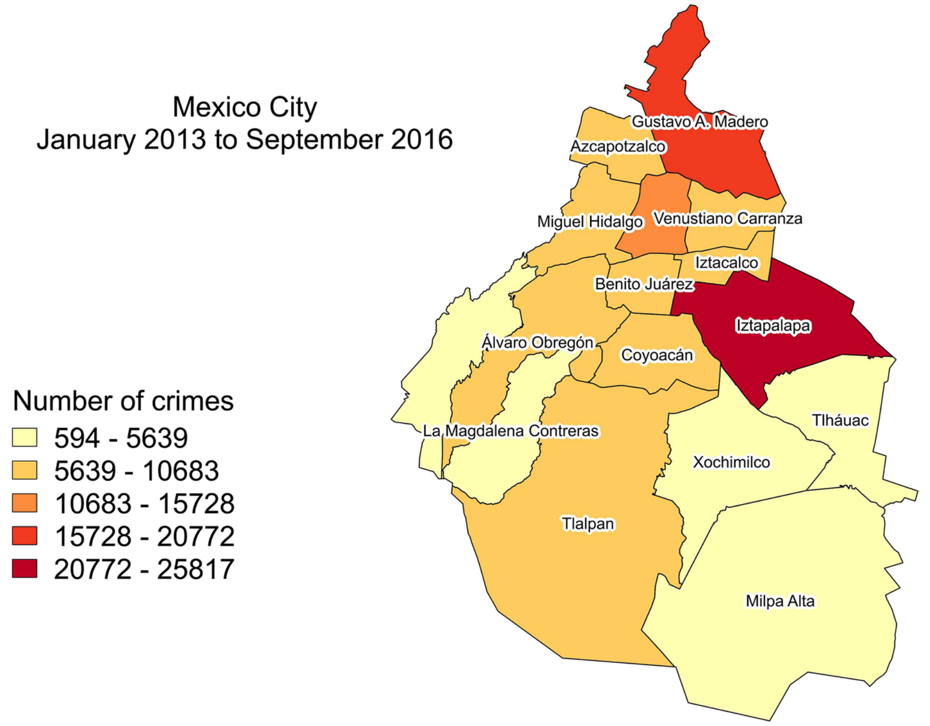

Fig. 1 Number of crimes per borough in Mexico City

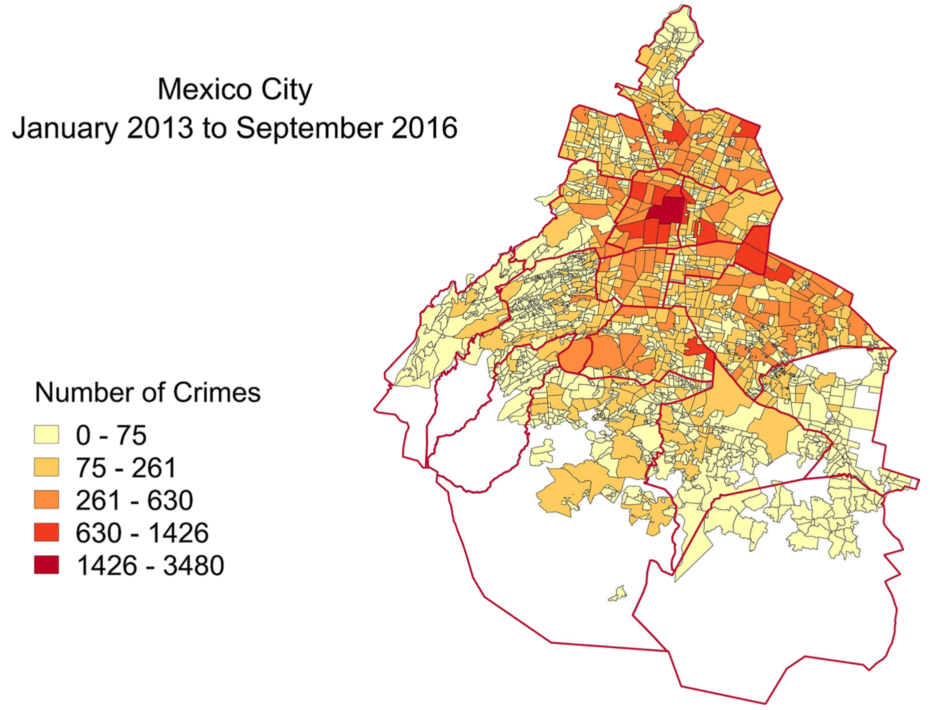

Fig. 2 Number of crimes per segmented neighborhoods in Mexico City

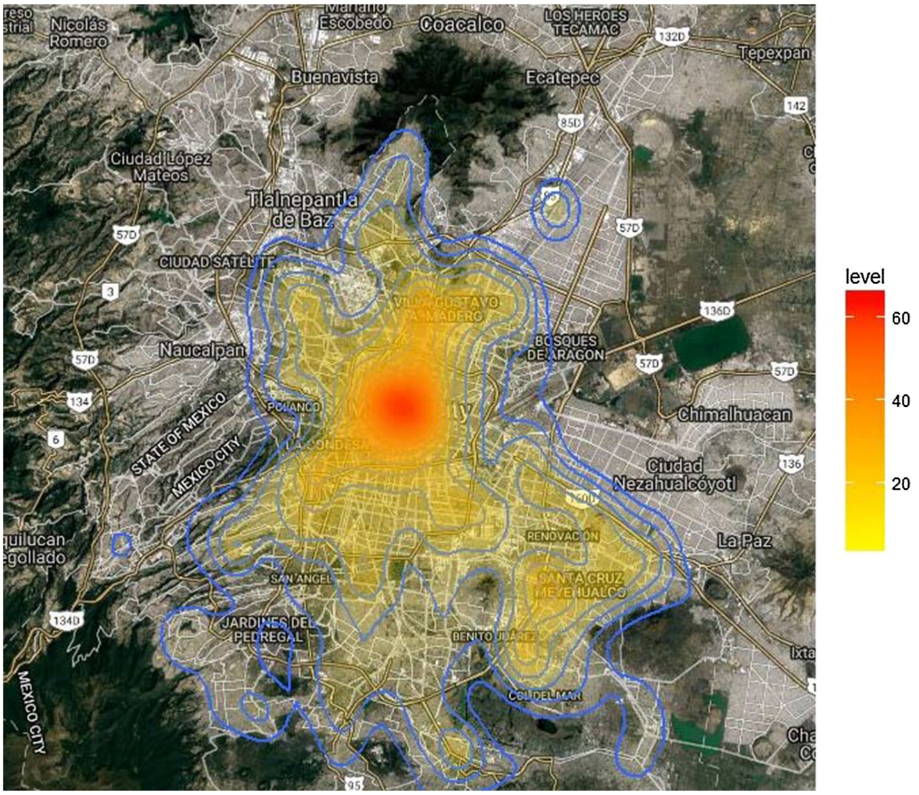

Fig. 3 Heatmap showing the relative intensity of crime frequency

note that some areas that apparently appear in a very low level crime region might be due to the lack of meaningful data.

Data and methods

Data sources for this study were as follows: official crime reports, Twitter and Google Trends. A more detailed account of data collection for each source is given in the following sections.

Official crime data

The Mexico City Police Department recorded 132,692 offenses from January 2013 to September 2016. This information related to crime reports was made accessible to the public and shared via [23]. However, the access to this information was suspended in 2016 due to the refusal of the Mexico City Police Department (SSP-CDMX) to provide updated crime reports. It was after the Mexico City elections in 2018 and the open data initiative proposed recently by the Mexico City government that it was possible to gain access to crime reports committed in Mexico City from 2016 to 2019. However, during this time window it is unclear which new criteria were used to arrange and classify crime reports in comparison to those collected from 2013 to 2016; for this reason, this manuscript cannot provide a comprehensive review of recent years.

Twitter data

Social media plays an important role in addressing the issue of sharing day to day events. In recent years, there has been an increasing interest in analysing human behavior [24, 25]. In this regard, we have collected publicly available tweets from September 2016 to April 2017 (time window) via the Twitter streaming API [26] which uses the push strategy for data retrieval. Once a request for information is made, the streaming API provides a continuous stream of updates with no further input from the user. Due to the Twitter streaming API time limitations in terms of historical data, this manuscript cannot provide a comprehensive review of the same time frame set by the Mexico City Police Department.

Given the nature of this case study, it is worth briefly discuss the ethical, legal, and social implications of using Twitter data to conduct research. Tweets that were collected through the public Twitter API are subject to the Twitter terms and conditions [27] and to the developer agreement and policy too [28]. The privacy policy used by Twitter indicates that users consent to the collection, transfer, manipulation, storage and disclosure of data that are public. Thus, this research inspected only those tweets that were public (i.e., no privacy settings were selected by the user) [29]. In order to comply with Twitter terms of service, data cannot be publicly shared. Interested future researchers may reproduce the experiments by following the procedure described in the following part of this manuscript. Non-anonymized data sets may be available upon request from the corresponding author.

By using a customized query string to filter tweets, we have gathered specific tweets that contain at least one of the following keywords in Spanish language: "inseguridad" (unsafe), "violencia" (violence), "robo" (robbery or burglary), "arma" (firearm), "violación" (rape and sexual assault), "asesinato" (murder), "crimen" (crime), "robo de auto partes" (auto parts theft) and "víctima" (victim). The process to retrieve data from Twitter was based on a "social explorer" aimed to collect and store filtered tweets in real-time [30].

Our sample consisted of 26,428 tweets that emerged during the observed time window. This data contains information such as: user ID, the screen name or alias, number of followers, date, the tweet itself, device used to post the tweet (source), the user-defined location, coordinates, age and gender.

It is important to note that approximately 1% of all tweets published on Twitter are geolocated. This is a very small portion of the tweets, and it is often necessary to use the profile information to determine the tweet's location [31]. However, most of the time the user profile information is not accurate and needs to be confirmed with a different method e.g., Internet Protocol address (IP address).

Similarly, even though Twitter data use the Global Positioning System (GPS) for generating geographical information with a relatively high resolution; with the worst-case accuracy of 7.8 m with 95% confidence, the precision of the data can be affected by atmospheric effects, receiver quality, sky blockage, and noise caused by weather or device factors. Therefore, two tweets sent from the same exact location could be reported as slightly different locations [32].

A process of data cleansing was carried out with the aim to detect, filter and remove corrupt or inaccurate records. We use a free and open source tool for working with messy and massive data [33]. Thus, it was possible to remove obvious errors e.g., nulled fields, empty sets and incomplete data. In addition, we discarded off-topic tweets in a semi-automated way by filtering only those tweets that were repeatedly posted [34]. Consequently, our corpus was reduced to 2572 (9.73% from the total of tweets) generated by 1494 users. The small size of the dataset meant that it was not possible to find more tweets related to crime in Mexico City. Due to this fact, it was necessary to suggest another source of data such as Google Trends.

Google Trends data

In the emerging digital society, search engines have become a significant tool for acquiring the latest relevant news about a target term. Google search engine is essential for a wide range of online searching tasks [35]. In particular, Google Trends is a free and accessible online portal that analyzes a portion of billions of daily Google searches, generating data on geographical and temporal patterns according to specified keywords [36]. Google Trends is the search volume for a given query relative to the total number of searches on Google on a scale of 0 to 100.

In order to investigate whether Google Trends can help to predict crime reports in Mexico city, we carried out a web search on this platform. The time limit of the Google Trends search matched exactly the time frame of official crime reports (from January 2013 to September 2016). Google Trends data provide much longer time series with easy geographical location. These two aspects were critical for our statistical analysis. The web search was based on a set of Spanish keywords as follows: "denunciar" (crime reporting), "denuncia" (crime report), "homicidio" (homicide), "lesiones" (injury), "robo" (theft), "asalto" (assault), "violación" (rape), "asesinato" (murder). It can be seen from this set of words that there is a high similarity between Twitter keywords and Google Trends keywords.

Results and discussion

Official crime reports

In order to understand how a selected range of crimes is experienced by residents in Mexico City a series of measurements were performed. First, a statistical analysis was used to compare between categories. Figure 4a depicts a comparison among the 13 categories defined by the Mexico City Police Department. In this case, theft of motor vehicle showed the highest frequency, around 53,411 reports followed by robbery passerby and robbery of business property with 34,524 and 13,398 reports respectively. A more detailed descriptive analysis is provided in Fig. 4b. These box plots indicate the degree of dispersion (spread) and skewness in official data.

Additionally, we plot the following time series graph provided in Fig. 5a. It can be seen from the graph that official data exhibit a linear downward trend in terms of theft of motor vehicle, robbery passerby and robbery of business property; the rest of the crime reports show no patterns or cycles over the same time period (from January 2013 to September 2016). A density plot that visualizes the distribution of the official data over the previously mentioned interval is presented in Fig. 5b. This chart shows smoother distributions. It is important to remark that the peaks of this plot display where the values are concentrated over the interval.

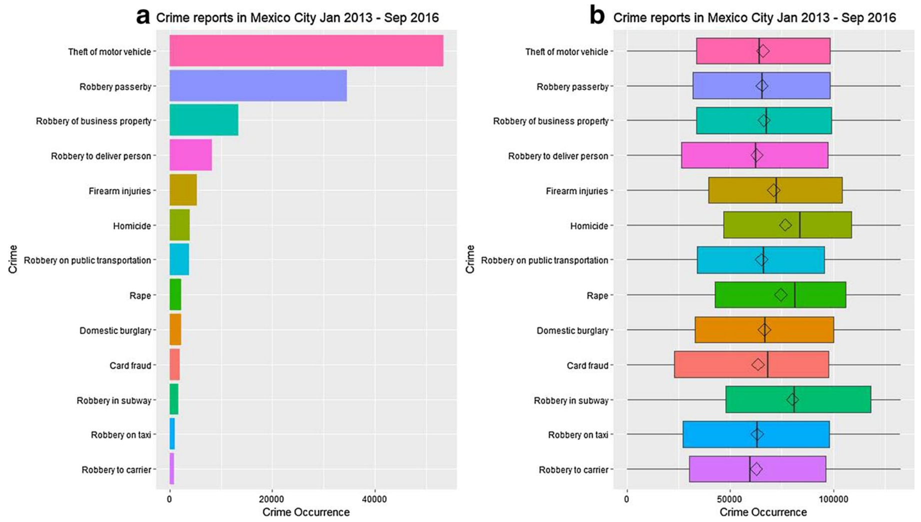

Fig. 4 Statistical analysis comparing among 13 categories of crime reports

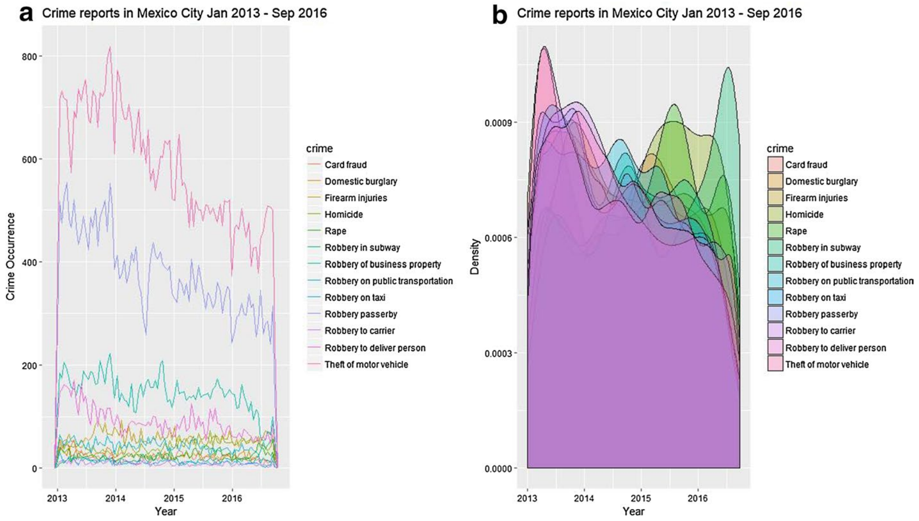

Fig. 5 Time series and density plot of crime reports in Mexico City

In order to compare the difference between crime reports from 2013 to 2016, four heat maps were generated where temporal data can be visualized in Figs. 6, 7, 8 and 9. These heat maps show the data value for each row and column standardized, so they all fit in the same range. Any pattern in the heat maps indicates a monthly association with the 13 crime categories. It can be seen that crime categories are converted into blue-green scale colors where the lowest value in the heat map is set to white, the highest value to a dark blue and mid-range values to light green with a corresponding transition (or gradient) between these extremes.

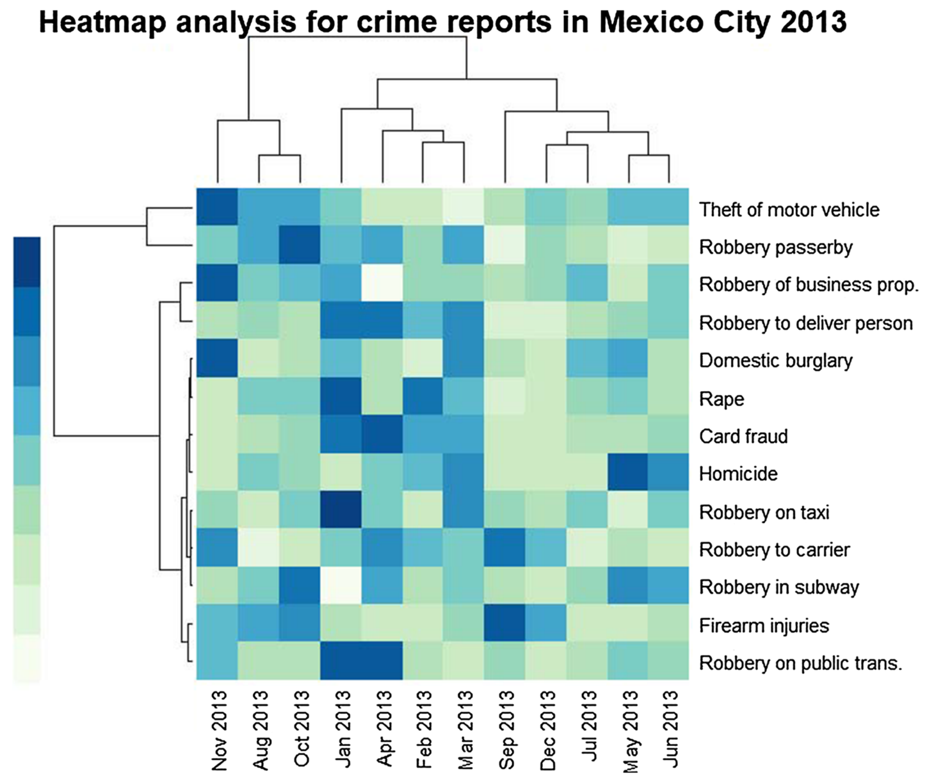

Fig. 6 Heat map analysis of crime reports in Mexico City: 2013

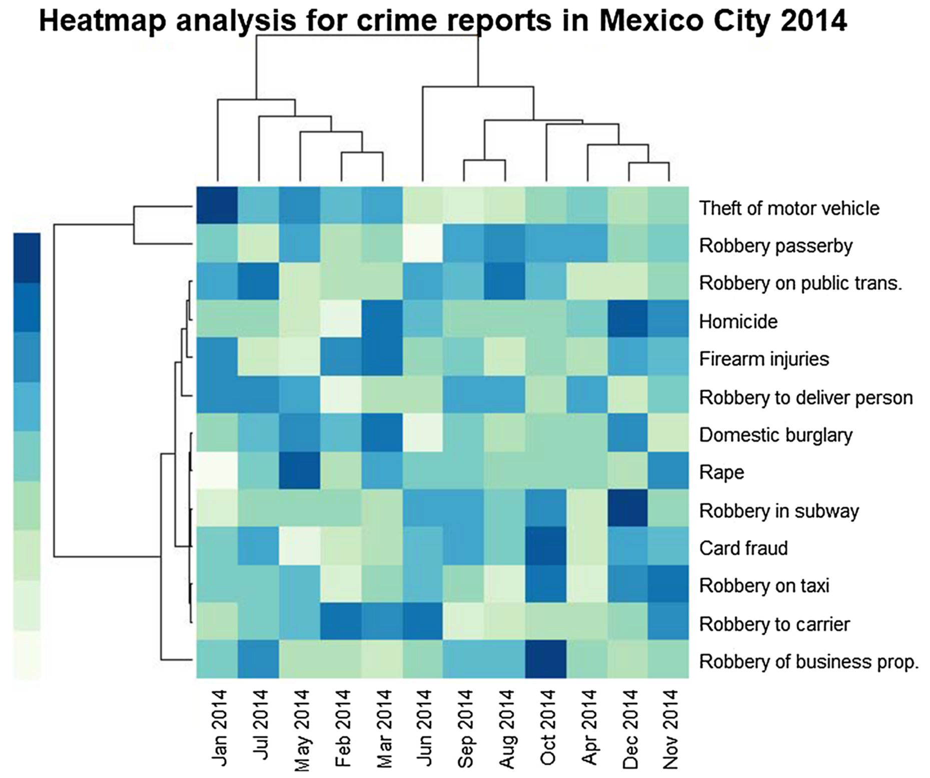

Fig. 7 Heat map analysis of crime reports in Mexico City: 2014

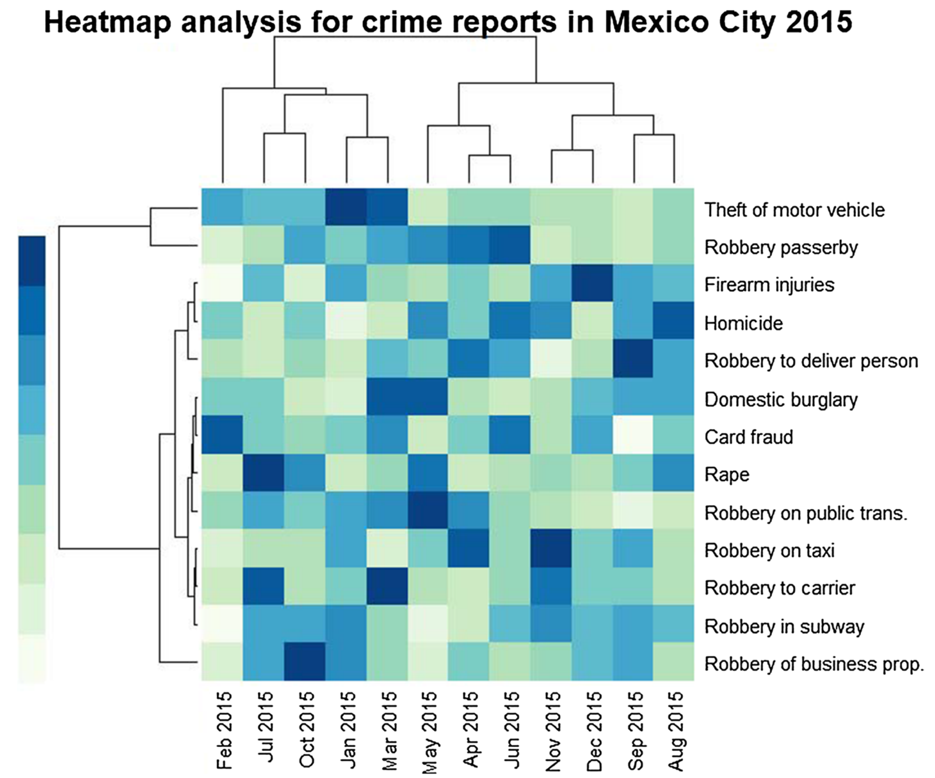

Fig. 8 Heat map analysis of crime reports in Mexico City: 2015

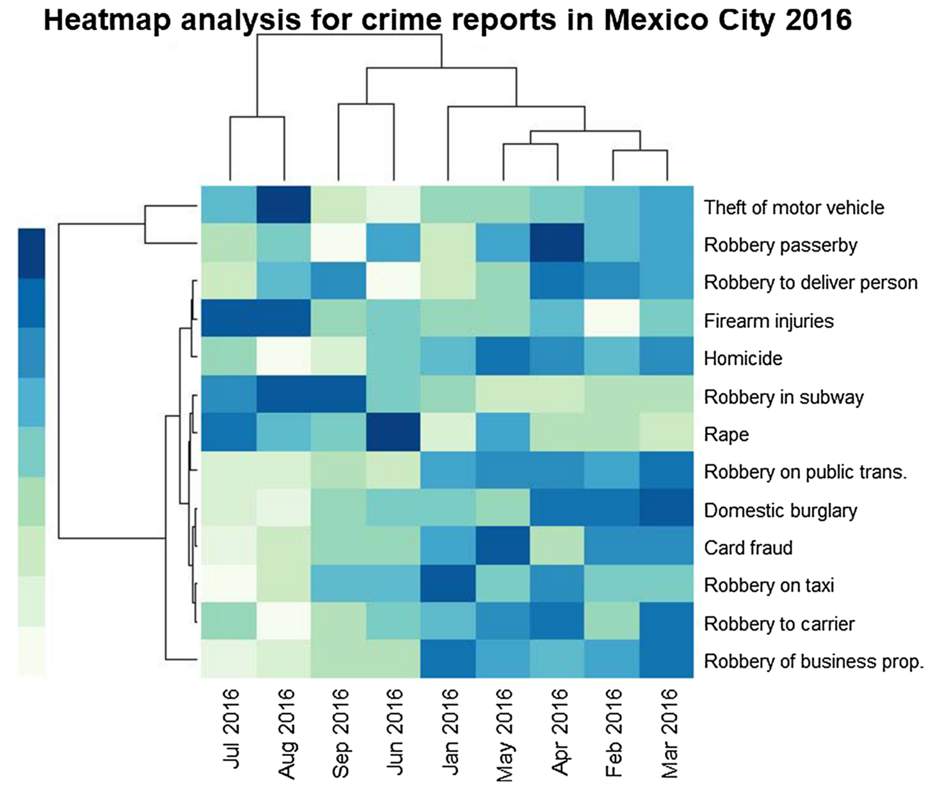

Fig. 9 Heat map analysis of crime reports in Mexico City: 2016

Twitter

Our sample of tweets was manually classified according to the official crime categories, this task was carried out empirically through visual inspection. Figure 10 presents a heatmap with the relative intensity of tweet occurrence, where a "hot" color indicates a high activity of tweets, and a "cold" color indicates a low activity of tweets. From this visual it can be observed that hot spots seem to be slightly consistent with the reported crime rates in Fig. 1.

We now present those accounts that reported significantly more activity according to our list of keywords. Nearly 20% of the tweets were generated through the Mexico City Police Department accounts: @SSP_CDMX and @PGJDF_CDMX. Interestingly, most users mentioned those two accounts in their tweet, as an unofficial complaint. These two accounts generally promote actions to prevent and reduce crime or criminal offenses. The rest of the accounts are associated to the news media and average Twitter users.

As described previously, we manually classified tweets into the official crime categories. Figure 11 provides an overview among the crime occurrence, the degree of dispersion and skewness. From the chart, it can be seen that robbery passerby has the highest number with around 1022 tweets followed by theft of motor vehicle and robbery of business property with 351 and 329 tweets respectively. It should be highlighted that these three categories are closely related to the top three crimes found in the official data.

Additionally, we plot time series graph and a density plot with the aim to illustrate crime activity posted on Twitter from September 2016 to April 2017. As shown in Fig. 12a, there are two categories that show the highest peaks: robbery passerby and theft of motor vehicle. In addition, this graph shows that the number of robberies of business property reached a peak during January 2017, this peak corresponds to the riots caused

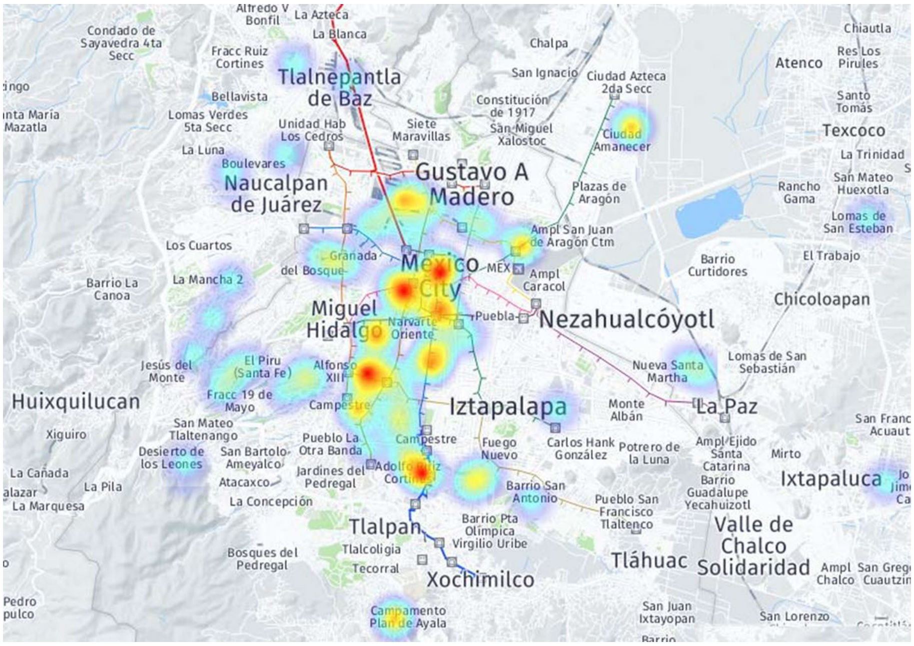

Fig. 10 Heatmap showing hot spots in terms of tweets

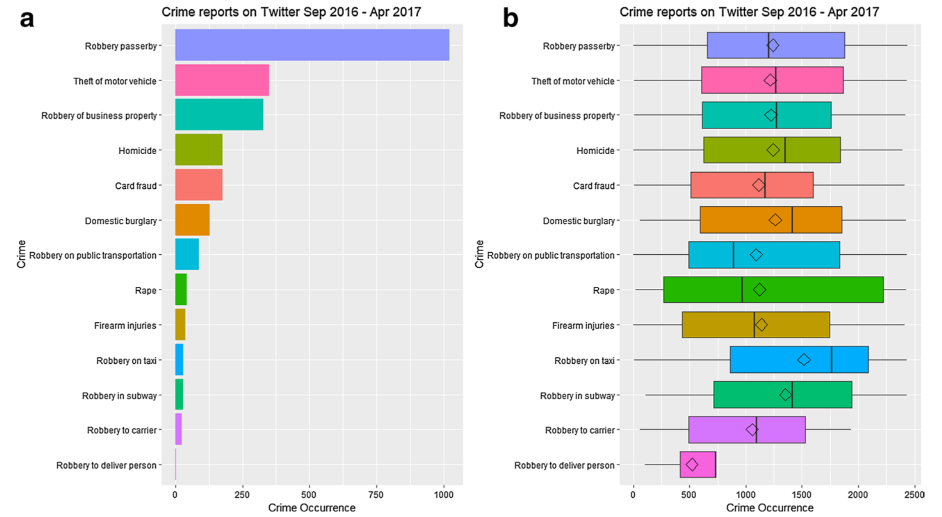

Fig. 11 Crime reports posted on Twitter

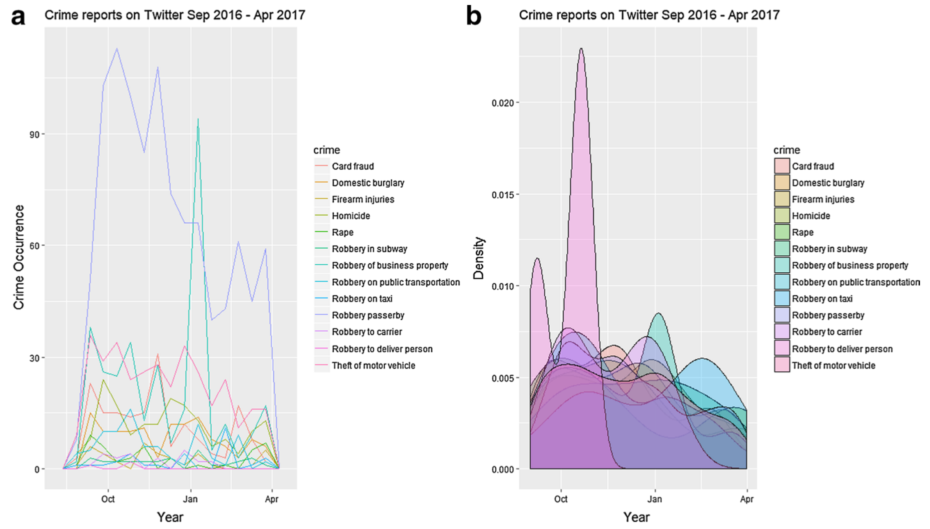

Fig. 12 Time series and density plot on Twitter

by the rising of fuel prices in Mexico. A density plot that visualizes a smoother distribution of the Twitter data is depicted in Fig. 12b.

The relative relationship between daily official crime reports and the proportion of collected tweets can be examined by the Pearson correlation coefficient. These values are depicted in Fig. 13. It can be seen from this array, that there is a moderate correlation between official data (y-axis) and our collected tweets (x-axis). The size of the boxes indicate their proximity to one (in absolute value) and the color shows the sign of the correlation coefficient blue for positive, red for negative. The degree of intensity of relationship in this case implies that a keyword in Twitter cannot fully describe or estimate any official crime report.

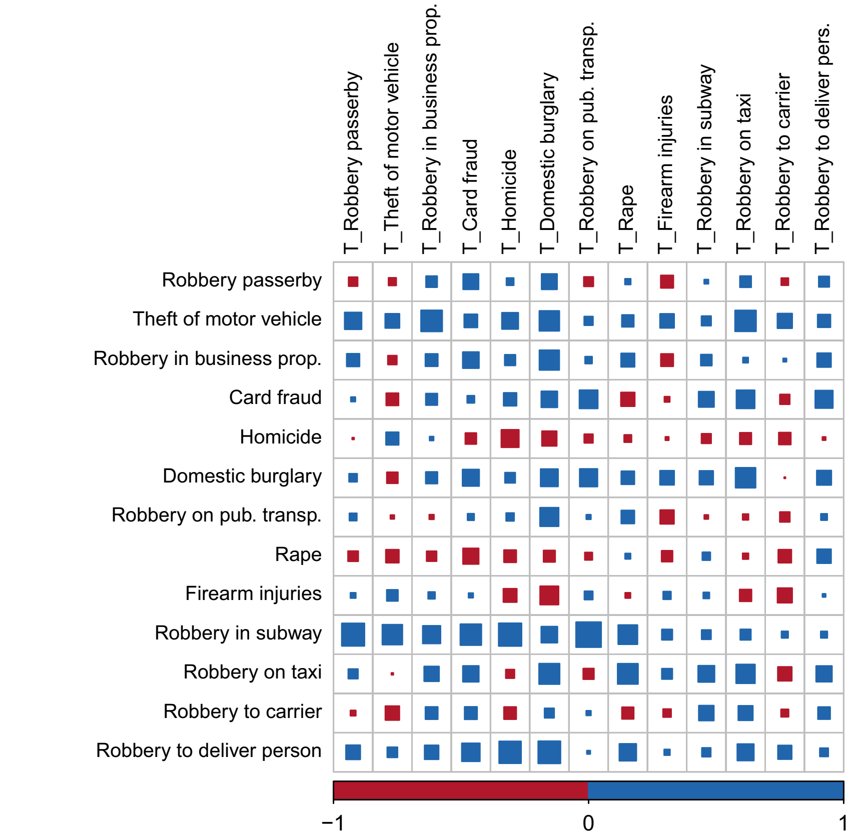

Fig. 13 Correlation between official crime data and tweets

Google Trends

Generally speaking, this section introduces Google Trends as informative predictors and proposes an online big-data-driven forecasting model for crime reports in Mexico City, then it investigates whether Google Trends can help to predict from an online big data perspective. Turning now to the experimental evidence on Google Trends, a Pearson correlation matrix in Fig. 14 was performed to determine the relationship between crime categories (y-axis) and Google Trends keywords (x-axis). The size of the boxes indicates its proximity to one (in absolute value) and the color shows the sign of the correlation coefficient blue for positive, red for negative. The degree of intensity of relationship in this case implies a limited correlation.

Due to the fact that ARIMA models are flexible and widely used in time series analysis, this method is suitable for time series of medium to long length [37]. Therefore, we applied an Autoregressive Integrated Moving Average model (ARIMA) on a weekly granularity to forecast crime in Mexico City. This model estimates the total number of crime reports from a day of the week to the corresponding one in the following week based on its previous observations. This model is explained as follows:

Let $Y_1, \ldots, Y_m$ be the weekly official data series of a specific crime, and let $X_t = X_{t,1}, \ldots, X_{t,r}$ be the number of Google Trends series for keywords or terms $i = 1, \ldots, r$. The considered autoregressive model for week t is

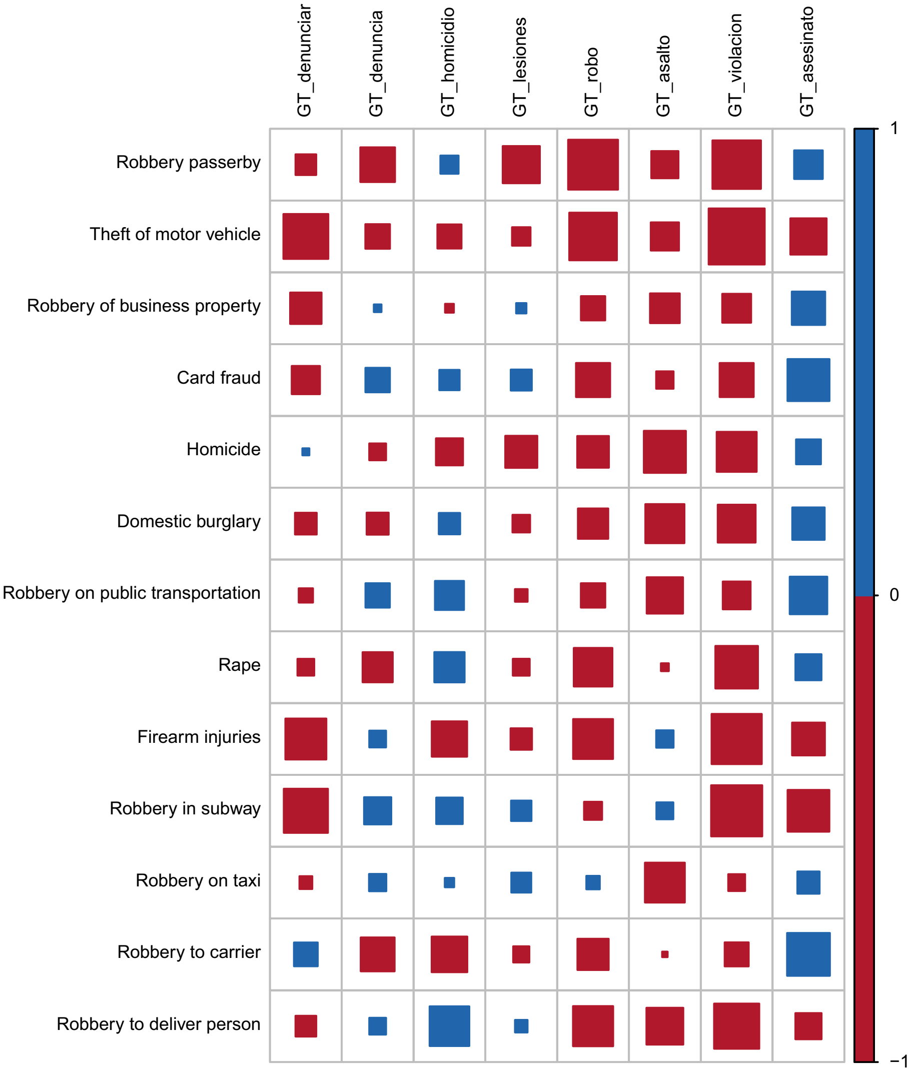

Fig. 14 Correlation between official crime and Google Trends data

$$Y_t^* = X_{t-1}'\boldsymbol{\beta} + \sum_{i=1}^{p} \phi_i Y_{t-i}^* + \sum_{j=1}^{q} \psi_j \nu_{t-j} + \nu_t, \qquad (1)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, $\nu_t \sim \mathrm{WN}(0, \sigma^2)$; $Y_t^* = (1-B)^d Y_t$, $p, d, q \in \mathbb{Z}^+$, $t = 1, \dots, m$; $\phi(\lambda) = 1 - \phi_1\lambda - \dots - \phi_p\lambda^p$; $\psi(\lambda) = 1 + \psi_1\lambda + \dots + \psi_q\lambda^q$, and $\phi(\lambda), \psi(\lambda) \neq 0, \forall|\lambda| \leq 1$.

The selection of p, d and q in this ARIMA model with covariates (also known as ARIMAX) was based on Akaike information criterion corrected for small sample sizes (AIC) and the remaining coefficients were estimated with least squares. The model selection and fit was carried out for each crime type and each week t, producing an adaptive procedure that is able to capture changes on the time series and its relations to the external variables $\boldsymbol{X}_t$. Therefore, model (1) is able to forecast the total number of reported crimes in the following week based on its previous weekly aggregated observations and Google Trends series.

Table 1 Percentage of parameter values selected for model (1)

Model	Order					
	0	1	2	3	4	Total
AR (p)	61.41	25.72	11.75	0.83	0.29	100
I (d)	48.3	51.7	0	0	0	100
MA (q)	39.91	52.15	7.28	0.66	0	100

Due to the fact that Google Trends series are highly associated with model (1), we have selected three specific series via the official crime data and the LASSO method. This method minimizes the residual sum of squares [38]. The Google Trends series selected correspond to "denuncia" (crime report),"robo" (theft) and "asesinato" (murder).

We fit adaptive models to obtain 1-week ahead predictions for all 13 official crime categories from week 11 to 196. Which means that, for a given week $t \in \{11, \ldots 196\}$ we fit the parameters p, d, q in model (1), with covariates $\boldsymbol{X}_{t-1,1}, \boldsymbol{X}_{t-1,2}, \boldsymbol{X}_{t-1,3}$ corresponding to the series, up to time $t - 1$, for "denuncia", "robo" and "asesinato", respectively. Then resulting p, d, q parameters, for all 2418 fitted models are summarized in Table 1, where it can be observed that 59.59% of all models do not show an autoregressive term and 51.08% of them present a moving average component of at least order 1.

Figure 15 Part 1: Homicide, Firearm injuries, Robbery in subway and Robbery on public transportation; Fig. 16 Part 2: Robbery on taxi, Domestic burglary, Card fraud and Robbery of business property; and Fig. 17 Part 3: Robbery to deliver person, Robbery passerby, Robbery to carrier, Theft of motor vehicle and Rape; depict the official crime data series with their corresponding 1 week forecasting for weeks 11 to 196. The black lines describe the official crime reports series, the red dots depict the 1 week point predictions and the shaded bands show the 95% predictive intervals. In general terms, these series are non-stationary and some unexpected changes are visible. In spite of this fluctuation, our adaptive fitted model is able to capture this type of trends and shifts in the series.

To assess the predictive power and sharpness of the obtained predictions, we get a measure of their length and the percentage of times that real observations fall within 80% and 95% predictive intervals. The improvement of the model calibration was evaluated through the comparison between the percentage of falls-in and their interval declared probability. As sharpness is related to interval precision, we have computed their average lengths (A.L.). It can be seen from the data in Table 2 that the coverage probability is close to the declared probability, except for "Homicide", "Domestic burglary" and "Robbery in subway".

In order to compare the interval lengths with different types of crime, we use average lengths that correspond to the mean of the interval length divided by the point that was forecast. Therefore, "Robbery on taxi" and "Robbery to carrier" were shown to have the largest intervals with falls-in above the 80% and 95% respectively. These results suggest that these predictive intervals might be reduced and still comply with 80% and 95% of coverage.

An ARIMA model without Google Trends covariates was developed with the aim to estimate predictions on the crime categories. In contrast to earlier findings, the coverage is

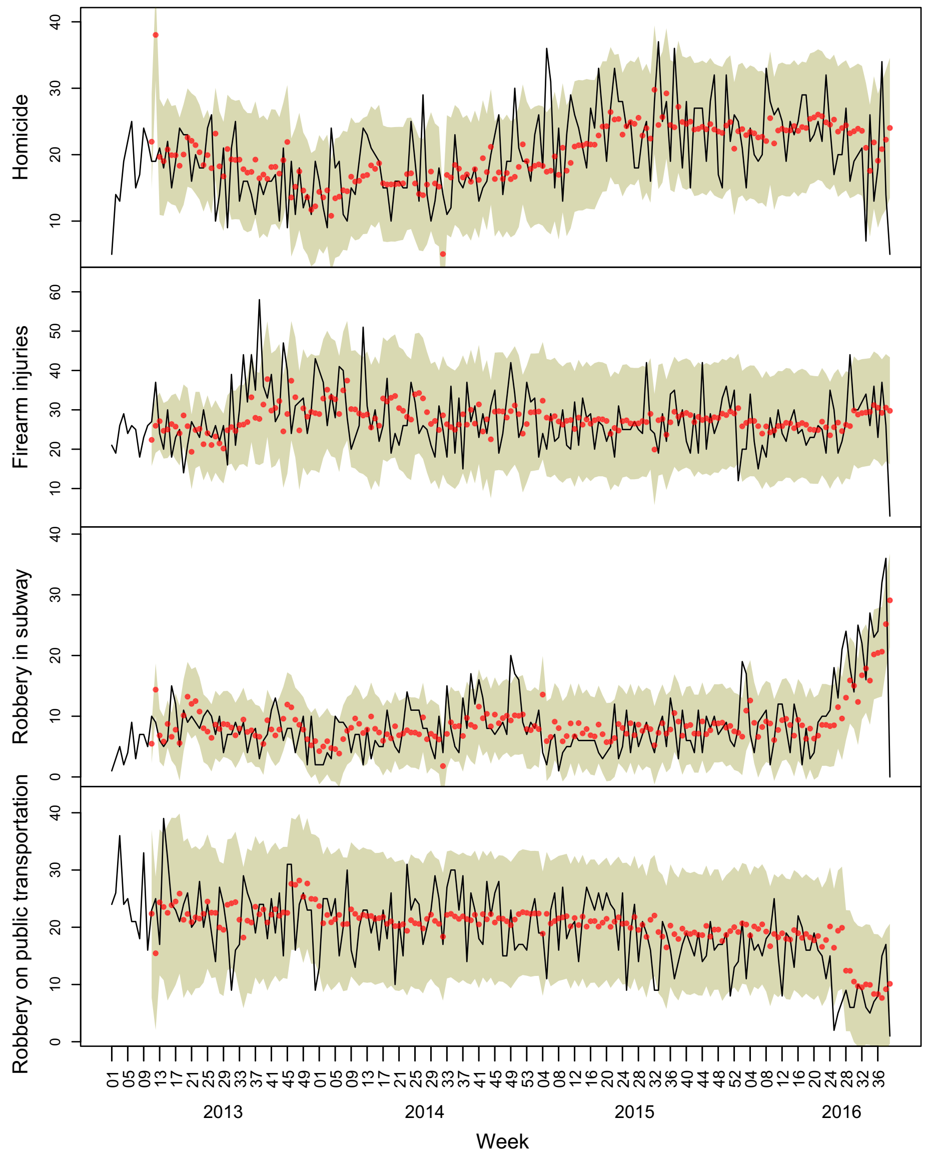

Fig. 15 One week adaptive forecasting for official crime dataset: part 1

closer to the declared 80% level for 9 out of 13 crimes and 95% level for 3 out of 13 crimes. On one hand, the averaged absolute errors for 80% and 95% coverage are 3.796 and 2.933 respectively for the previous ARIMAX model. On the other hand, for this ARIMA model the averaged absolute errors for 80% and 95% coverage are 4.507 and 1.873 respectively.

Discussion

In this study we performed an extensive analysis of the crime patterns in Mexico City. The results of this research reflect differences in identifying official crime reports and collected data from social media sources such as Twitter and Google Trends. In this regard, it was found that theft of motor vehicle (including auto parts theft), robbery

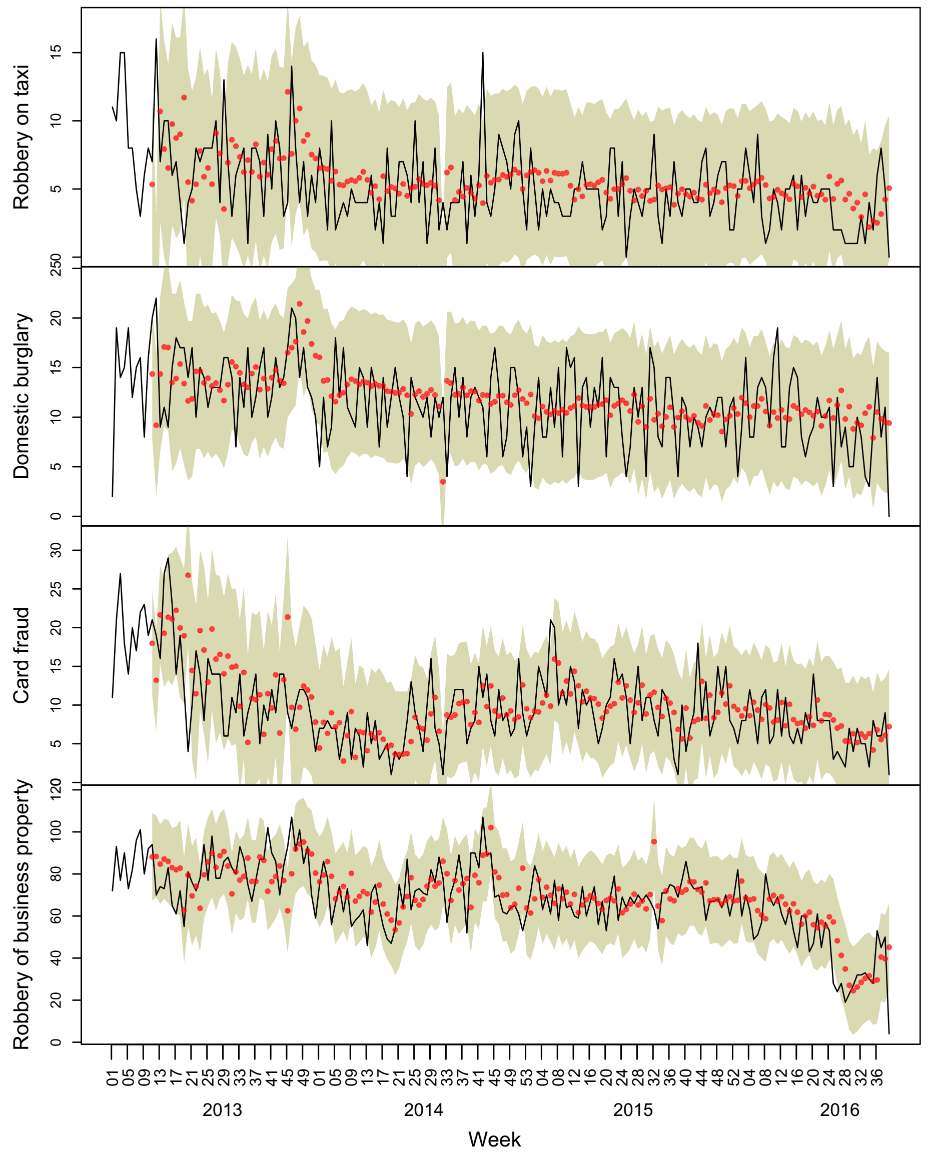

Fig. 16 One week adaptive forecasting for official crime dataset: part 2

passerby and robbery of business property were the most common crimes recorded by the local police and social media.

Despite the fact that the pairwise correlation between official crime reports and data series obtained from Twitter or Google Trends was almost negligible, as is depicted in Figs. 13, 14. These findings, while preliminary, suggest that the use of online data sources could be statistically significant and a suitable approach that might help to improve and strength the official crime reports in terms of forecasting models.

This finding broadly supports the work of other studies e.g., in epidemiology there is an extensive number of studies dedicated to these ideas [39–43]. In addition, few studies have investigated crime reports or rate prediction [15, 44–46].

This study provided the relative intensity of crime occurrence through descriptive maps. This analysis confirms that crime is clustered in certain areas and boroughs e.g.,

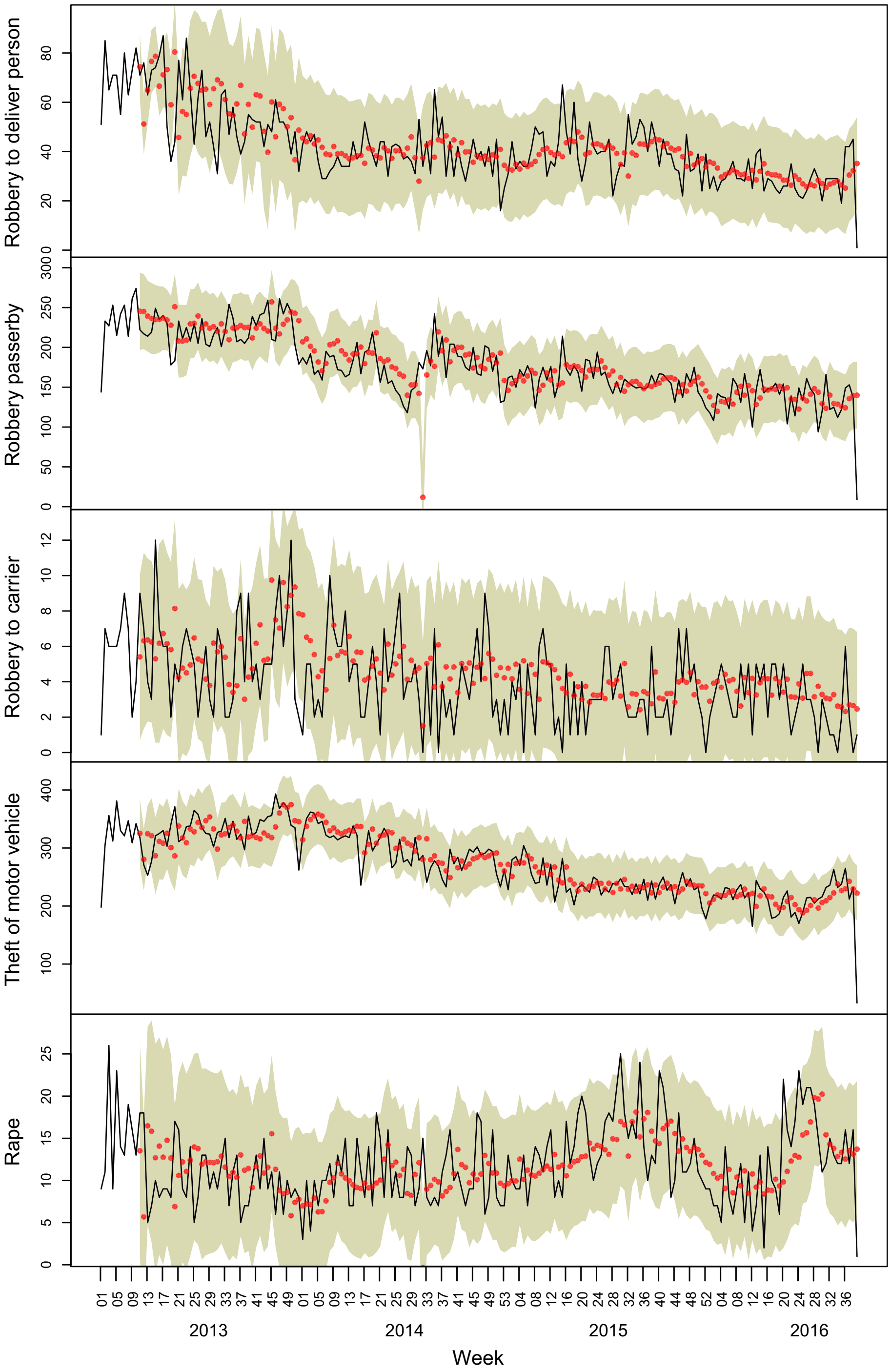

Fig. 17 One week adaptive forecasting for official crime dataset: part 3

the inner city. However, at this point we are not able to determine the origin and the causes of these clusters.

It is important to note, that data taken from social media are not necessarily representative and it must be considered as a mere observational information that

Table 2 Percentage of fall-in for 80% and 95% predictive intervals and their average length (A.L.)

Crimes	80%		95%	
	% Fall-in	A.L.	% Fall-in	A.L.
Robbery passerby	80.65	0.36	92.47	0.54
Theft of motor vehicle	82.80	0.25	93.01	0.39
Robbery of business property	77.96	0.44	89.78	0.67
Card fraud	81.18	1.24	94.09	1.90
Homicide	70.43	0.62	87.63	0.95
Domestic burglary	70.97	0.82	91.40	1.26
Robbery on public transportation	80.11	0.77	94.62	1.18
Rape	81.18	0.72	93.55	1.10
Firearm injuries	77.96	0.66	90.32	1.00
Robbery in subway	66.67	1.07	84.41	1.64
Robbery on taxi	83.87	1.37	95.70	2.10
Robbery to carrier	79.03	1.56	95.16	2.39
Robbery to deliver person	85.48	0.73	95.16	1.11

enhances the official crime reports. Nevertheless, this case of study suggest the potential of using social media data to understand how people report or share a criminal offense on social media.

A linear predictive model was used to assess the performance of Google Trends in predicting crime rates. The proposed adaptive models explained earlier are able to detect the shifting behavior of the reported crimes. Here, we assume that the LASSO selected Google Trends series include meaningful information that remains through the period. However, the suggested prediction can be adaptively extended to the covariates. Thus, parameters *p*, *d*, *q* and Google Trends series were chosen weekly to be used as covariates. As a result of this, forecasting was more reliable and accurate to online information phenomenon such as self-excitement [47].

It may be said that models combining Google data with demographic information generally have lower error rates than models using only demographic data. Thus, external sources can be potentially helpful to enhance statistical predictions. Our preliminary results provide an early response to improve efforts in order to prevent crime. It should be noted that social media data is highly noisy which makes these findings less generalizable to our predictive model. However, our model adds to the rapidly expanding field of Big Data, an interesting approach to strength official data with online sources.

Conclusions

The analysis of crime reports undertaken here has extended our knowledge about crime patterns and antisocial behavior in Mexico City. Moreover, this study has shown that Twitter may be useful in understanding the spatio-temporal patterns of crime data. Additionally, our predictive model confirmed that Google Trends works as a valid predictor to anticipate the impact of crime. To the best of our knowledge, this is the first study that systematically reviews and examines online sources of data related to crime reports in Mexico City. Preliminary results of this research show that social media data (Twitter and Google Trends) can be integrated to enhance Big Data-driven models for predicting crime rates. These results could be used by the Mexico City Police Department to develop tailored strategies to tackle crime and gradually build community trust and confidence on the local police.

The analysis we present here provides evidence that data generated through Twitter and Google Trends can be used to generate estimates of the crime occurrence. In particular, we illustrate how the choice of time period and geographic area analysed can impact the outcome of such an analysis. Finally, we found that more targeted work will need to be done to fully understand these data sources and the potential opportunities for developing more accurate forecasting models.

Abbreviations

CDMX: Mexico City; API: Application Programming Interface; GPS: Global Positioning System; user ID: user Identification Device; ARIMA: Autoregressive Integrated Moving Average; ARIMAX: Autoregressive Integrated Moving Average with covariates; AIC: Akaike information criterion; LASSO: least absolute shrinkage and selection operator; A.L.: average length; CCTV: closed circuit television.

Acknowledgements

This research has been supported in part by "Premio a la investigación interdisciplinaria". C. A. Piña-García was partially supported by SNI membership 69310.

Authors' contributions

CAPG and LRR conceived of the presented idea and developed the theory and performed the computations. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The Mexico City Police Department have released the yearly official crime reports and the datasets analysed during the current study are available from the following websites: https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/custom/ or https://hoyodecrimen.com/en/ Twitter data have been collected through the public Twitter API (https://developer.twitter.com/en.html). Therefore, to comply with Twitter terms of service, data cannot be publicly shared. Interested future researchers may reproduce the experiments by following the procedure described in the paper. Non-anonymized data sets may be available upon request from Dr. C. A. Piña-García (cpina@uv.mx).

Ethics approval and consent to participate

Not applicable.

Consent for publication

The authors consent for publication.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Laboratorio para el Análisis de Información Generada a través de las Redes Sociales en Internet (LARSI), Centro de Estudios de Opinión y Análisis, Universidad Veracruzana, Xalapa, Mexico. [2] Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Mexico.

Received: 29 March 2019 Accepted: 10 July 2019
Published online: 22 July 2019

References

1. Chadee D, Ng Ying NK, Chadee M, Heath L. Fear of crime: the influence of general fear, risk, and time perspective. J Interpers Violence. 2019;34(6):1224–46.
2. U.S.DS: Overseas Security Advisory Council. 2018. https://www.osac.gov/pages/ContentReportDetails.aspx?cid=17114. Accessed 09 June 2018.
3. PGJ: Procuraduria General de Justicia de la CDMX. 2018. http://www.pgj.cdmx.gob.mx/. Accessed 01 Jan 2019.
4. SSP: Secretaría de Seguridad Pública de la CDMX. 2018. http://www.ssp.cdmx.gob.mx/. Accessed 01 Jan 2019.
5. Vilalta CJ. Fear of crime in public transport: research in Mexico City. Crime Prev Community Saf. 2011;13(3):171–86.
6. CNDH: Comisión Nacional de los Derechos Humanos - México. 2019. http://www.cndh.org.mx/. Accessed 01 Jan 2019.

7. Davies TP, Bishop SR. Modelling patterns of burglary on street networks. Crime Sci. 2013;2(1):10.
8. Rosser G, Davies T, Bowers KJ, Johnson SD, Cheng T. Predictive crime mapping: arbitrary grids or street networks? J Quant Criminol. 2017;33(3):569–94.
9. Davies T, Johnson SD. Examining the relationship between road structure and burglary risk via quantitative network analysis. J Quant Criminol. 2015;31(3):481–507.
10. Oliveira M, Bastos-Filho C, Menezes R. The scaling of crime concentration in cities. PLoS ONE. 2017;12(8):0183110.
11. Espinal-Enríquez J, Larralde H. Analysis of Mexico's narco-war network (2007–2011). PLoS ONE. 2015;10(5):0126503.
12. González F. Drug trafficking organizations and local economic activity in Mexico. PLoS ONE. 2015;10(9):0137319.
13. Wang M, Gerber MS. Using twitter for next-place prediction, with an application to crime prediction. In: 2015 IEEE symposium series on computational intelligence. IEEE. 2015. pp. 941–8.
14. Malleson N, Andresen MA. Spatio-temporal crime hotspots and the ambient population. Crime Sci. 2015;4(1):10.
15. Aghababaei S, Makrehchi M. Mining social media content for crime prediction. In: 2016 IEEE/WIC/ACM international conference on web intelligence (WI). IEEE. 2016. pp. 526–31.
16. Chen X, Cho Y, Jang SY. Crime prediction using twitter sentiment and weather. In: 2015 systems and information engineering design symposium. IEEE. 2015. pp. 63–8.
17. Flores RD. Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. Am J Sociol. 2017;123(2):333–84.
18. Yadav N, Kumar A, Bhatnagar R, Verma VK. City crime mapping using machine learning techniques. In: International conference on advanced machine learning technologies and applications. Springer. 2019. pp. 656–68.
19. Yang D, Heaney T, Tonon A, Wang L, Cudré-Mauroux P. Crimetelescope: crime hotspot prediction based on urban and social media data fusion. World Wide Web. 2018;21(5):1323–47.
20. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could google trends be used to predict methamphetamine-related crime? an analysis of search volume data in Switzerland, Germany, and Austria. PLoS ONE. 2016;11(11):0166566.
21. D'Avanzo E, Pilato G, Lytras M. Using twitter sentiment and emotions analysis of google trends for decisions making. Program. 2017;51(3):322–50.
22. Sampson RJ, Raudenbush SW, Earls F. Neighborhoods and violent crime. In: Community health equity: a Chicago Reader. 2019. p. 282.
23. de Crimen H. Crime in Mexico City. 2019. https://hoyodecrimen.com/en/. Accessed 01 Jan 2019.
24. Budiharto W, Meiliana M. Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. J Big Data. 2018;5(1):51.
25. AlMahmoud H, AlKhalifa S. Tsim: a system for discovering similar users on twitter. J Big Data. 2018;5(1):39.
26. Twitter: Developer Twitter API. 2019. https://developer.twitter.com/en/docs. Accessed 01 Jan 2019.
27. Twitter: Terms of Service. 2019. https://twitter.com/en/tos. Accessed 01 Jan 2019.
28. Twitter: Developer Agreement and Policy-Twitter Developers. 2019. https://developer.twitter.com/en/developer-terms/agreement-and-policy. Accessed 01 Jan 2019.
29. McIver DJ, Hawkins JB, Chunara R, Chatterjee AK, Bhandari A, Fitzgerald TP, Jain SH, Brownstein JS. Characterizing sleep issues using twitter. J Med Internet Res. 2015;17(6):e140.
30. Piña-García C, Gershenson C, Siqueiros-García JM. Towards a standard sampling methodology on online social networks: collecting global trends on twitter. Appl Netw Sci. 2016;1(1):3.
31. Zheng X, Han J, Sun A. A survey of location prediction on twitter. IEEE Trans Knowl Data Eng. 2018;30(9):1652–71.
32. Wang Q, Phillips NE, Small ML, Sampson RJ. Urban mobility and neighborhood isolation in America's 50 largest cities. Proc Natl Acad Sci. 2018;115(30):7735–40.
33. Refine O. Open Refine. 2019. http://openrefine.org/. Accessed 01 Jan 2019.
34. Ham K. Openrefine (version 2.5). http://openrefine.org.free, open-source tool for cleaning and transforming data. J Med Libr Assoc. 2013;101(3):233.
35. Yu L, Zhao Y, Tang L, Yang Z. Online big data-driven oil consumption forecasting with google trends. Int J Forecast. 2019;35(1):213–23.
36. Cervellin G, Comelli I, Lippi G. Is google trends a reliable tool for digital epidemiology? insights from different clinical settings. J Epidemiol Glob Health. 2017;7(3):185–9.
37. Albayrak AS. Arima forecasting of primary energy production and consumption in turkey: 1923–2006. Enerji piyasa ve düzenleme. 2010;1(1):24–50.
38. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267–88.
39. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012.
40. De Angelis D, Presanis AM, Birrell PJ, Tomba GS, House T. Four key challenges in infectious disease modelling using data from multiple sources. Epidemics. 2015;10:83–7.
41. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol. 2015;11(10):1004513.
42. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep. 2015;5:12760.
43. Xu Q, Gel YR, Ramirez-Ramirez LL, Nezafati K, Zhang Q, Tsui K-L. Forecasting influenza in Hong kong with google search queries and statistical model fusion. PLoS ONE. 2017;12(5):0176690.
44. Bolla RA. Crime pattern detection using online social media. Master's thesis, Missouri University of Science and Technology. 2014.
45. Domdouzis K, Akhgar B, Andrews S, Gibson H, Hirsch L. A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations. J Syst Inf Technol. 2016;18(4):364–82.
46. Ristea A, Leitner M. Integration of social media in spatial crime analysis and prediction models for events. In: AGILE PhD School. 2017.
47. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. Science. 2014;343(6176):1203–5.