Big Data
Volume 3 Number 3, 2015

DOI: 10.1089/big.2014.0054

ORIGINAL ARTICLE

Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics

Andrey Bogomolov,[1] Bruno Lepri,[2,*] Jacopo Staiano,[3] Emmanuel Letouzé,[4,5] Nuria Oliver,[6] Fabio Pianesi,[2] and Alex Pentland[7]

Abstract

The wealth of information provided by real-time streams of data has paved the way for life-changing technological advancements, improving the quality of life of people in many ways, from facilitating knowledge exchange to self-understanding and self-monitoring. Moreover, the analysis of anonymized and aggregated large-scale human behavioral data offers new possibilities to understand global patterns of human behavior and helps decision makers tackle problems of societal importance. In this article, we highlight the potential societal benefits derived from big data applications with a focus on citizen safety and crime prevention. First, we introduce the emergent new research area of *big data for social good*. Next, we detail a case study tackling the problem of crime hotspot classification, that is, the classification of which areas in a city are more likely to witness crimes based on past data. In the proposed approach we use demographic information along with human mobility characteristics as derived from anonymized and aggregated mobile network data. The hypothesis that aggregated human behavioral data captured from the mobile network infrastructure, in combination with basic demographic information, can be used to predict crime is supported by our findings. Our models, built on and evaluated against real crime data from London, obtain accuracy of almost 70% when classifying whether a specific area in the city will be a crime hotspot or not in the following month.

Key words: big data analytics; data mining; predictive analytics; crime

Introduction

The transition of data from being a scarce resource to a massive and real-time processed stream is rapidly changing the world we live in, challenging and often subverting long-lasting paradigms in a broad range of domains.[1] Finance, economics, politics, journalism, medicine, biology, and physics, to name a few, have been disrupted by the existence of large amounts of data. The almost universal adoption of the mobile phone and the exponential growth of Internet services has led to the existence of unprecedented amounts of data about human behavior.

In this context, it is important to differentiate between two use cases when it comes to *big data*: (1) *personal data* applications, where data of (anonymized) individuals are analyzed at the individual level to build computational models of each person to, for example, provide personalized services or adapt the interaction (in this use case, privacy, security, transparency, control, and accountability are key elements that need to be taken into account), and (2) *aggregate data* applications, where aggregated and anonymized data of individuals are analyzed collectively to be able to make inferences about large-scale human behavior. In this scenario, as long as the level of aggregation is sufficiently large, no data can be traced back to any individual and hence there are minimal—if any—privacy concerns.

[1]*University of Trento, Trento, Italy.*
[2]*Fondazione Bruno Kessler, Trento, Italy.*
[3]*Sorbonne Universités, Paris, France.*
[4]*University of California–Berkeley, Berkeley, California.*
[5]*Data-Pop Alliance, New York, New York.*
[6]*Telefonica Research, Barcelona, Spain.*
[7]*MIT Media Lab, Cambridge, Massachusetts.*

*Address correspondence to: *Bruno Lepri, Mobile and Social Computing Lab, Fondazione Bruno Kessler, via Sommarive 18, I-38123 Povo, Trento, Italy*, E-mail: lepri@fbk.eu

The work presented in this article falls in the context of this second use case, and in particular, within the emergent research area of *big data for social good*, that is, the prospect of leveraging big (aggregate) data to positively affect policy and society.

Although still "*in its intellectual and operational infancy*,"[2] the area of big data for social good has gone through a rapid phase of expansion and maturation in a short period of time—driven by key research studies on mapping the propagation of diseases such as malaria[3] and H1N1 flu,[4] monitoring socioeconomic deprivation,[5] predicting human emergency behavior,[6] detecting the impact of natural disasters such as floods,[7] and inferring pollution emissions of vehicles,[8] but also driven by nonacademic institutions (e.g., United Nations Global Pulse, Flowminder.org, Data-Pop Alliance, and DataKind) and initiatives (e.g., Orange Data for Development, Telefonica Datathon for Social Good, Telecom Italia Big Data Challenge, and Chicago Data Science for Social Good Fellowship). A recent report published by UN Global Pulse[9] discussed the challenges and opportunities of using big data for societal challenges and proposed a three-tier taxonomy of uses: "real-time awareness," "early warning," and "real-time feedback." A subsequent article on the specific case of big data for conflict prevention distinguished its "descriptive" (i.e., maps), "predictive" (i.e., either proxying or forecasting), and "prescriptive" (i.e., the realm of causal inference) functions.[10] Whereas interest in the latter function is poised to grow, most applications have relied on the first two—and perhaps most visibly on the second.

This has notably long been true for counterterrorism, intelligence, and law-enforcement activities—with "predictive policing" systems as near-perfect examples. Critics of these approaches have pointed to their inability to tackle root causes and the risks of profiling and harassment they may create, while questioning their efficiency.[11] Others have argued that curbing crime preemptively may have lasting structural impacts on communities plagued by violence.

Crime has not yet been widely covered in the big data for social good literature, apart from a few examples.[12–15] However, it provides fertile ground to advance our common understanding of crime and to validate the power of *place-based crime models* built from anonymized and aggregated human behavioral data with limited to no privacy risks.

In this article, we propose and evaluate a big data approach to the problem of crime hotspot classification, that is, the identification, based on past data, of geographic locations that are likely to become scene of a crime. In particular, we combine demographics with anonymized and aggregated *people dynamics* features, derived from mobile network activity, in order to classify whether specific locations are more or less likely to become crime hotspots in the near future. Note that none of our data sources can be traced back to make inferences about individuals.

Crime Hotspots Classification

Crime is a well-known social problem affecting the quality of life and the economic development of a society. Several works have shown that crime tends to be associated with slower economic growth at both the national level[16] and the local level, such as cities and metropolitan areas.[17] Dating back to the beginning of the 20th century, studies have focused on the behavioral evolution of criminals and its relations with specific characteristics of the neighborhoods in which they grew up, lived, and acted. Existing works tend to mainly explore relationships between criminal activity and socioeconomic variables such as education,[18] ethnicity,[19] income level,[20] and unemployment.[20]

Urbanists and architects have also investigated the relationships between people dynamics, urban environment, and crime.[21,22] Urban activist Jane Jacobs[21] has emphasized *natural surveillance* as a key deterrent for crime: as people are moving around an area, they will be "eyes on the street" able to observe what is going on around them. Hence, Jacobs suggests that high diversity among the population and high number of visitors contribute to the safety of a given area and lead to less crime. On the contrary, Newman's theory[22] argues that a high mix of people creates the anonymity needed for crime. Thus, according to the latter, low population diversity, low visitors ratio, and a high ratio of residents are the features contributing to an area's safety. Several studies have tried to shed light onto these conflicting theories. Felson and Clarke[23] have proposed the routine activity theory, which investigates how specific situations and variations in lifestyle affect the opportunities for crime. Specifically, they found that some places such as bars and pubs attract crime.

Criminologists have also started to investigate in detail significant concentrations of crime at microlevels of geography, regardless of the specific unit of analysis.[24] Research has shown that in what are generally seen as good parts of town there are often streets with strong crime concentrations, and in what are often defined as bad neighborhoods, there are locations relatively

free of crime.[24] In 2008, criminologist David Weisburd proposed to switch the popular people-centric paradigm of police practices to a place-centric paradigm.[25]

Based on these findings, we adopt a *place-centric* and *data-driven* approach: specifically we investigate the power of *people dynamics*—derived from a combination of mobile network activity and demographic information—to determine whether a *specific geographic area* is likely to become a scene of the crime.

Analyzed Datasets

For our case study we exploit datasets provided during a public competition—the Datathon for Social Good—organized by Telefonica Digital, The Open Data Institute, and the MIT Human Dynamics Group. This Datathon took place in the context of the Campus Party Europe 2013 at the O2 Arena in London in September 2013.

Participants were provided access to the following data, among others:

- *Anonymized and aggregated human behavioral and demographics data* computed from mobile network activity and demographics information in the London Metropolitan Area. We shall refer to this dataset as the Smartsteps dataset, because it was derived from Telefonica's Smartsteps product.
- *Geolocalized open data*, a collection of openly available datasets with varying temporal granularity. This includes reported criminal cases, residential property sales, transportation, weather, and London borough profiles related to homelessness, households, housing market, local government finance, and societal well-being (a total of 68 metrics).

We turn now to describing the specific datasets that we used to classify crime hotspots.

Criminal cases dataset

The criminal cases dataset includes the geolocation of all reported crimes in the United Kingdom but does not specify their exact date, just the month and year. The data provided in the public competition included the criminal cases for December 2012 and January 2013.

The dataset includes the crime ID, the month and year when the crime was committed, its location (longitude, latitude, and address where the crime took place), the police department involved, the lower layer super output area (LSOA) code, the LSOA name, and the crime type out of 11 possible types (antisocial behavior, burglary, violent crime, shoplifting, etc.).

LSOAs are small geographical areas (mean population of 1,500 and minimum population threshold of 1,000) defined by the United Kingdom Office for National Statistics following the 2001 census. Their aim here is to define areas, based on population levels, whose boundaries would not change over time.

Smartsteps dataset

The Smartsteps dataset consists of a geographic division of the London Metropolitan Area into cells whose precise location (latitude and longitude) and surface area were provided. Note that the actual shape of the cell was not provided. In total, there were 124,119 cells. For each of the Smartsteps cells, a variety of demographic variables were provided, computed every hour for a 3-week period, from December 9th to 15th, 2012, and from December 23rd, 2012, to January 5th, 2013. In particular,

1. *Footfall*, or the estimated number of people within each cell. This estimation is derived from the mobile network activity by aggregating every hour the total number of unique phone calls in each cell tower, mapping the cell tower coverage areas to the Smartsteps cells, and extrapolating to the general population—by taking into account the market share of the network in each cell location.
2. An estimation of *gender, age, and home/work/visitor group splits*. That is, for each Smartsteps cell and for each hour, the dataset contains an estimation of how many people are in the cell; the percentage of these people who are at home, at work, or just visiting the cell; and their gender and age splits in the following brackets: 0–20 years, 21–30 years, 31–40 years, etc., as shown in Table 1. This information is not directly available from the activity in the phone network infrastructure but was provided by GFK, a market research firm.

London borough profiles dataset

The London borough profiles dataset is an official open dataset containing 68 different metrics about the population of a particular geographic area. The spatial granularity of the borough profiles data is at the LSOA level.

The information includes statistics about the population, households (census), demographics (proportion of population aged 0–15 in 2011, proportion of working age population in 2011, proportion of population

Table 1. Smartsteps data provided by the challenge organizers

Type	*Data*
Origin based	Total no. of people No. of residents No. of workers No. of visitors
Gender based	No. of males No. of females
Age based	No. of people aged up to 20 No. of people aged 21–30 No. of people aged 31–40 No. of people aged 41–50 No. of people aged 51–60 No. of people aged over 60

All the demographic variables refer to 1 h intervals and to each Smartsteps cell.

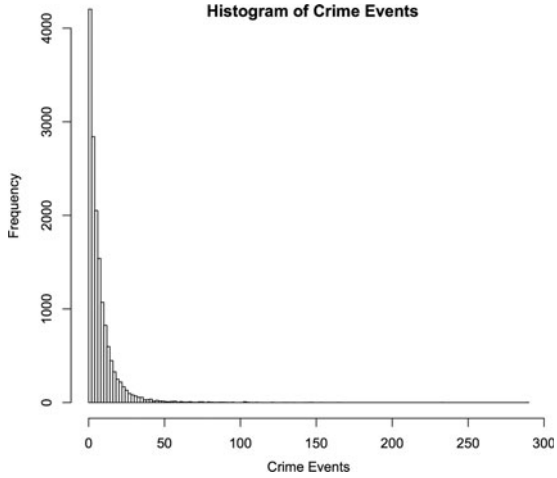

FIG. 1. Spatial distribution of crime events.

aged 65 or over in 2011, etc.), migrant population (e.g., proportion of largest, second largest, and third largest migrant population by country of birth in 2011), ethnicity (e.g., proportion of population from black, Asian, and minority ethnic groups), language (e.g., proportion of people aged 3+ whose main language is not English), employment (e.g., female, male, and total employment rate in 2012), NEET (Not in Education, Employment, and Training) people, benefits (e.g., proportion of the working age population who claim out work benefits in 2012), qualifications (e.g., proportion of the working age population with no qualifications in 2012), earnings (e.g., male, female, and general gross annual pay in 2012), volunteering, jobs density, business survival, crime, fires, house prices, new homes, tenure, greenspace, recycling, carbon emissions, cars (e.g., number of cars and number of cars per household in 2011), indices of multiple deprivation, General Certificate of Secondary Education (GCSE) results, children in out-of-work families, life expectancy, teenage conceptions, happiness levels, political control (e.g., proportion of seats won by Labour, LibDem, and Conservatives), and election turnout.

Classifying Crime Levels

We cast the problem of crime hotspot classification as a binary classification task. For each Smartsteps cell, we classify whether it will show a *high* or *low* crime level in the next month. In order to do this, we use Smartsteps features computed on December data and crime ground-truth observations from January. The formulation of the problem as a binary classification task is driven by several motivations. First, an important advantage of dichotomizing the ground-truth variable is that it greatly simplifies the presentation of the results so to be easily understandable to a wide audience. This is the reason why, often, dichotomization is used in criminology studies—where one of the goals is to present results to policy makers and police departments.[26,27] Second, given the fixed and finite resources available, policy makers and police departments are mainly interested in having a simple tool to decide where to allocate "more" versus "less" resources while leaving the quantification of these resources to the decision maker. Finally, dichotomizing a continuous variable, and in particular dichotomizing using the median split, is statistically convenient when dealing with highly skewed distributions.[28] As depicted in Figure 1, the majority of Smartsteps cells in our dataset have few crimes (e.g., only 1 crime event), while in a small proportion of the cells a high number of crimes is observed. The spatial distribution of the criminal cases for the month of January is summarized in Table 2. Given the high skewness of the distribution (skewness = 5.88, kurtosis = 72.5, mean = 8.2, median = 5), we split the criminal dataset with respect to its median into two classes: *a low crime* (class "0") when the number of crimes in the given cell was less or equal to the median, and *a high crime* (class "1") when the number of crimes

Table 2. Number of crime cases in January

Min.	*Q1*	*Median*	*Mean*	*Q3*	*Max.*
1	2	5	8.2	10	289

in a given cell was larger than the median. Hence, a *crime hotspot* is a cell with a number of crimes strictly higher than the median value of crimes for that particular month (hence, labeled as high crime). Following the empirical distribution, the two resulting classes are approximately balanced (53.15% for the high-crime class).

The separation among training and testing sets is done spatially, with 80% of the cells used for training and 20% of the cells used for testing. It is worth to note that no classification of past labels based on future Smartsteps data is possible since Smartsteps features are computed from December data, while crime ground-truth labels are computed exclusively for January. In the following subsections we provide details of the experimental setup that we followed.

Referencing geotagged data to Smartsteps cells

As the Smartsteps cell IDs, the borough profiles, and the crime event locations are not spatially linked in the provided datasets and we did not have access to the actual shape of the Smartsteps cells, we first georeferenced each crime event by identifying the Smartsteps cell centroid closest to the location of the crime. We carried out a similar process for the borough profiles dataset. As a result, each crime event and the borough profile information were linked to one of the Smartsteps cells. In order to accomplish this, we implemented the approximation Algorithm 1 (see Fig. 2). Accounting for the curvature of the earth, we introduced Algorithm 2 (see Fig. 3) to calculate the direct spatial distance, given the Federation Aeronautique Internationale Earth model, such that the Earth is treated as a three-dimensional ellipse, defined by two radii, a major axis (the radius at the equator), and a minor axis (the radius at the poles). The major axis is set to a constant equal to 6371.009 km.

Feature extraction

Diversity and *regularity* have been shown to be important in the characterization of different facets of human behavior. In particular, the concept of entropy[29] has been applied to assess the socioeconomic characteristics of places and cities,[5] the predictability of mobility,[30] and spending patterns.[31] Hence, for each Smartsteps variable (see Table 1), we computed the mathematical functions that characterize its distribution and information theoretic properties, for example, mean, median, standard deviation, min and max values, and Shannon entropy.[29]

Algorithm 1: Approximating closest Telefonica's Output Area Centroid for each Crime Event

Function: getClosestCentroid
Input: $\{lat_{crime}, long_{crime}\} \in \mathbb{R}$.
Output: $\{c_{id}, d_{crime}, r_c\}$.

begin
  Initialization:
  $R \longleftarrow 6371.009$
  $c_{id} \longleftarrow \emptyset$, $d_{crime} \longleftarrow \emptyset$, $r_c \longleftarrow \emptyset$, $D \longleftarrow \emptyset$
  foreach $i \in$ *Telefonica Output Area Centroids* **do**
    $D_{crime} \longleftarrow$ *directDistance*$(lat_c, long_c, lat_{crime}, long_{crime})$
  end
  Sort ascending D
  Select firt row from D $\{c_{id}, c_{id}, r_c\}$
  return $\{c_{id}, d_{crime}, r_c\}$
  end

FIG. 2. Algorithm 1: Approximating closest Telefonica's output area centroid for each crime event.

Furthermore, in order to be able to also account for temporal relationships within the Smartsteps data, the same computations described above were repeated on sliding windows of variable length (1 h, 4 h, and 1 day), producing second-order features that help reduce computational complexity and the feature space itself, while preserving useful data properties.

Algorithm 2: Esimating Direct Distance

Function: directDistance
Input: $\{lat_1, long_1, lat_2, long_2\} \in \mathbb{R}$.
Output: $d \in \mathbb{R}$.

begin
  Initialization:
  $R \longleftarrow 6371.009$
  $lat_d \longleftarrow \emptyset$, $long_d \longleftarrow \emptyset$, $a \longleftarrow \emptyset$, $c \longleftarrow \emptyset$
  $lat_d \longleftarrow \frac{\pi * (lat_2 - lat_1)}{180}$
  $long_d \longleftarrow \frac{\pi * (long_2 - long_1)}{180}$
  $a \longleftarrow \sin(\frac{lat_d}{2}) * \sin(\frac{lat_d}{2}) + \cos(\frac{\pi * lat_1}{180}) * \cos(\frac{\pi * lat_2}{180}) * \sin(\frac{long_d}{2}) * \sin(\frac{long_d}{2})$
  $c \longleftarrow 2 * \arctan(\frac{\sqrt{1-a}}{\sqrt{a}})$
  return $R * c$
end

FIG. 3. Algorithm 2: Estimating direct distance.

Conversely, no data preprocessing was needed for the London borough profiles. Hence, we used the original 68 London borough profile features.

Feature selection

One of our goals is to provide a comparison between our approach—based on Smartsteps data—against a traditional one—based on borough profiles data. Hence, we decided to limit the number of features used by our model to 68, to match the maximum number of borough profiles variables that we were granted access to. Moreover, the limitation of the number of features reduces training times and enhances generalization performance by reducing the risk of overfitting.[32]

To this end, a feature subset selection step was performed following a bootstrap aggregating (bagging) procedure and using exclusively data from the training set. Bagging is a machine learning procedure, wherein predictors are constructed using bootstrapped samples from the training set and then aggregated to form a "bagged predictor." Each bootstrapped sample is formed using a dropout strategy, leaving out 1/3 of the training examples. These left-out examples are used to form accurate estimates of important measurements for local optimization decisions (e.g., to give better estimates of node probabilities and node error rates in decision trees).

The metric used for feature ranking was the mean decrease in the Gini coefficient of inequality.[33] This choice was motivated because it outperformed other metrics such as mutual information, information gain, and chi-square statistic.[33]

The Gini coefficient ranges between 0, expressing perfect equality (all dimensions have the same predictive power) and 1, expressing maximal inequality in predictive power. The feature with maximum mean decrease in Gini coefficient is expected to have the maximum influence in minimizing the out-of-the-bag error, namely, the misclassification error rate that is estimated on the dropped-out samples during the bagging procedure. It is known in the literature that minimizing the out-of-the-bag error results in maximizing common performance metrics used to evaluate models.[34]

The top 20 features selected by the model are included in Table 3.

Model building

Classification was performed by means of Random Forests (RF) ensemble classifiers.[35] We chose RF because they satisfy the max-margin property, they do not require parameter tuning, and, more importantly, they do not require the specification of a feature-space, as support vector machines (SVMs) do through the kernels. Moreover, RF are one of the most accurate learning algorithms available.[36,37] We ran the same experiments described below also by using SVMs with linear and RBF kernels and we obtained less stable and less accurate results on both training and testing sets. Hence, we report the performance results only for the best model, which was based on RF.

Table 3. Top 20 selected features ranked by mean in decrease accuracy

Base feature	*Temporal resolution*	*1st order*	*2nd order*	*0*	*1*	*Mean decrease accuracy*	*Mean decrease gini*
Age >60	Daily	Entropy.empirical	Entropy.empirical	4.48	5.43	9.02	18.75
At home	Daily	Mean	SD	3.20	7.60	8.91	27.13
Age <20	Daily	SD	Entropy.empirical	5.69	3.97	8.85	16.88
Age <20	Daily	Mean	Entropy.empirical	3.09	5.88	8.85	17.26
Age <20	Daily	Mean	SD	4.50	5.27	8.65	16.03
At home	Daily	Min	Entropy.empirical	6.39	2.32	8.61	15.99
At home	Daily	SD	SD	3.22	8.58	8.60	45.82
At home	Daily	SD	Mean	3.35	5.83	8.57	24.93
Age >60	Daily	Entropy.empirical	SD	4.62	4.95	8.56	20.45
At home	Daily	SD	Median	5.41	5.04	8.50	26.48
Age 31–40	Daily	Entropy.empirical	Max	2.33	5.79	8.44	16.24
Age 31–40	Daily	Min	SD	6.81	4.06	8.31	36.52
At home	Daily	Min	SD	4.36	6.85	8.29	34.26
At home	Daily	SD	Max	4.13	6.87	8.27	34.89
At home	Monthly	Max	—	3.92	5.42	8.26	29.86
At home	Monthly	SD	—	4.43	4.17	8.21	39.70
Age 51–60	Daily	Entropy.empirical	Entropy.empirical	4.74	4.11	8.13	16.64
Age <20	Daily	SD	SD	3.67	5.88	8.12	16.86
At home	Daily	Entropy.empirical	Entropy.empirical	5.13	4.82	8.08	18.55
At home	Daily	Max	SD	2.83	6.29	8.07	26.85

Decision trees are an intuitive method to tackle classification and regression problems. In the case of binary classification, a tree assigns features by creating a control structure on a feature middle point for a decision of splitting either left or right through nodes of the tree depending on the value of a given point of the variable. A binary tree, by definition, ensures that each case of independent variable is assigned to a unique terminal node. The value of the terminal node is a predicted outcome and defines the classification decision. That means that the decision rule is a path down the tree to its terminal node. The decision boundary is estimated by an ensemble set of decision rules. The RF algorithm produces a combination of trees, such that each one is dependent on the values of a random vector sampled independently with the same distribution for all the classification trees in the forest.[35] We took advantage of the well-known performance improvements that are obtained by growing an ensemble of trees and voting for the most frequent class. Random vectors were generated before the growth of each tree in the ensemble, and a random selection without replacement was performed.[35]

Experimental Results

In this section we report the experimental results obtained by the RF trained on different subsets of the selected features and always tested on the test set, which was not used during the training phase in any way.

The performance metrics used to evaluate our approach are (1) accuracy, (2) F1 score, the harmonic mean between precision, and recall, and (3) area under the ROC (AUC) score.[38,39]

In order to understand the value added by the Smartsteps data, we compared the performance of the RF[35] using all features (Smarsteps + borough) with (1) a baseline majority classifier, which always returns the majority class ("high crime") as its prediction (accuracy = 53.15%), and with two additional models trained with (2) only the subset of selected features derived from the borough profiles dataset (borough-only), and (3) only the subset of selected features derived from the Smartsteps dataset (Smartsteps-only).

Table 4 reports accuracy, F1 score, and the AUC metric for each of the models. First, the Smartsteps + borough and the Smartsteps models significantly outperform the baseline majority classifier, with an increase of about 15% of accuracy.

Table 4. Metrics comparison

Model	*Accuracy*	*Accuracy, 95% CI*	*F1 score*	*AUC*
Smartsteps + borough profiles	68.83	0.67, 0.70	68.52	0.63
Smartsteps	68.04	0.66, 0.69	67.66	0.63
Borough profiles	62.18	0.60, 0.64	61.72	0.58
Majority Classifiers (Baseline)	53.15	0.53, 0.53	0	0.50

AUC, area under the ROC.

Interestingly, the addition of the borough profiles features does not yield any significant improvement to the Smartsteps-only model (Smartsteps + borough model accuracy = 68.83 vs. Smartsteps-only model accuracy = 68.04). Moreover, the borough-only model yields a competitive but significantly lower accuracy than the Smartsteps model: 62.18%, more than 6% lower than the accuracies obtained with the Smartsteps-only model (68.37%) and with the Smartsteps + borough models (68.83%) while using the same number of variables.

In Table 4 we also report the F1 score for each model. This metric is the harmonic mean of the precision (the number of correct positive results divided by the number of all positive results) and the recall (the number of correct positive results divided by the number of positive results that should have been returned), where an F1 score reaches its best value at 1 and worst score at 0.[38,39]

Looking more in detail the performances of the different models on the "high crime" class, we focus on the *true-positive rate*, namely, the proportion of actual high-crime cells that are correctly identified, and on the *true-negative rate*, namely, the proportion of actual low-crime cells that are correctly identified. The Smartsteps-only and Smartsteps + borough models obtain a good true-positive rate performance of 74.20% and 73.90%, respectively. Instead, the only-borough model reaches a true-positive rate of 68.81%, over 5% less than the models based on Smartsteps data.

When looking at the true-negative rate performances, all the models obtain a worse performance: 63.07% for Smartsteps + borough, 61.06% for only-Smartsteps, and 54.66% for only-borough. Interestingly, our approach obtains a true-positive rate about 10% higher than the true-negative rate. Thus, our approach is performing better in correctly identifying high-crime cells. It is worth emphasizing that, in our scenario, it is more relevant to obtain good results on the "high crime" class: in fact, mistakenly assigning "high crime" to a cell is less dangerous, from a social

Table 5. Confusion matrix for the only-borough model

	Actual	
	0	***1***
Predicted		
0	786	509
1	652	1123

policy perspective, than erroneously classifying it as "low crime" when indeed it is "high crime." As the results show, our proposed approach brings significant advantages for the task of hotspot prediction.

For detailed analyses, Tables 5–7 report the confusion matrices of the only-borough model, the only-Smartsteps model, and the Smartsteps + borough model, respectively.

Finally, a visual comparison of the ROC curves for each of the models is provided in Figure 4.

Discussion and Implications

The results discussed in the previous section show that human behavioral and demographic data (at a daily and monthly scale) significantly improves the prediction accuracy when compared to using rich statistical data about a borough's population (households census, demographics, ethnicity, employment, etc.). The borough profiles data provides a fairly detailed view of the living conditions of a particular area in a city, yet this data is expensive and time-consuming to collect. Hence, this type of data is typically updated with low frequency (e.g., every few years), making difficult the observation of potential changes.

Human behavioral data derived from mobile network activity combined with demographics, though less comprehensive than borough profiles, provides significantly finer temporal and spatial resolution.

Next, we focus on the most relevant predictors of crime level taking a look at the top 20 variables in our model, which are sorted by their mean reduction in accuracy in Table 3. The Smartsteps features have more predictive power than official statistics coming from borough profiles: no features listed in the top 20 are obtained using borough profiles.

Table 6. Confusion matrix for the only-Smartsteps model

	Actual	
	0	***1***
Predicted		
0	878	421
1	560	1211

Table 7. Confusion matrix for the Smartsteps + borough model

	Actual	
	0	***1***
Predicted		
0	907	426
1	531	1206

Moreover, higher-level features extracted over a sequence of days from variables encoding the daily dynamics have more predictive power than features extracted on a monthly basis. This finding points out at the importance of capturing the temporal dynamics of a geographical area in order to predict its levels of crime.

Interestingly, features derived from the percentage of people in a certain cell who are at home both at a daily and monthly basis seem to be of extreme importance. In fact, 11 of the top 20 features are related to the *at home* variable. Newman's approach of "defensible space"[22] postulates the relevance of a high number of residents in an area to reduce crime. The predictive power of home variables seems to confirm their relevance. However, we found positive associations between the home variables and crime. Hence, our findings do *not* support Newman's thesis,[22] suggesting

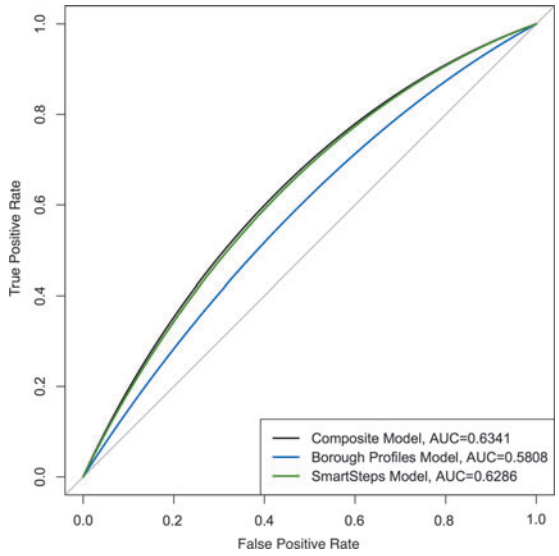

FIG. 4. ROC curves for Smartsteps + borough, only-Smartsteps, and only-borough profiles.

that an increased ratio of residents is linked to less crime and higher urban safety. Similar results were found in recent work done by Traunmueller et al.[15] In their work, the researchers focus only on testing some hypotheses about people dynamics and crime using correlational analyses between footfall counts recorded by the mobile network activity and crime activities.

It is also interesting to note the role played by the unpredictability of the variables, captured by Shannon entropy features.[29] The entropy-based features in fact seem useful for predicting the crime level of places (8 features out of the top 20 are entropy-based features). In our study, the Shannon entropy captures the predictable structure of a place in terms of the types of people that are in that area over the course of a day. A place with high entropy would have a lot of variety in the types of people visiting it on a daily basis, whereas a place with low entropy would be characterized by regular patterns over time. In this case, the daily unpredictability in patterns related to different age groups, different use (home vs. work), and different genders seems to be a good predictor for the crime level in a given area. In line with our results, Traunmueller et al.[15] found significant negative correlations between areas with higher age diversity and crime. Both our findings and those of Traunmueller et al.[15] support Jacobs's theory[21] of natural surveillance that high diversity of functions in an area and high diversity of people (gender diversity and age diversity) act as "eyes on the street" decreasing the number of crimes.

Interestingly, Eagle et al.[5] found that Shannon entropy used to capture the social and spatial diversity of communication ties within an individual's social network was strongly and positively correlated with economic development. Hence, high-diversity areas seem to emerge as more safe and more economically developed.

Our proposed approach could have clear practical implications by informing police departments and city governments on how and where to invest their efforts and on how to react to criminal events with quicker response times. From a proactive perspective, the ability to predict the safety of a geographical area may provide information on explanatory variables that can be used to identify underlying causes of these crime occurrence areas and hence enable officers to intervene in very narrowly defined geographic areas.

The distinctive characteristic of our approach lies in the use of features computed from aggregated and anonymized mobile network activity data in combination with some demographic information. Previous research efforts in criminology have tackled similar problems using background historical knowledge about crime events in specific areas,[40,41] criminals' profiling,[42] or wide description of areas using socioeconomic and demographic indicators.[43] Our findings provide evidence that aggregated and anonymized data collected by the mobile infrastructure, combined with demographic information, contains relevant information to describe a geographical area in order to classify its crime level.

The first advantage of our approach is its predictive ability. Our method classifies crime level using variables that capture the dynamics and characteristics of the demographics and nature of a place rather than only making extrapolations from previous crime histories. Operationally, this means that the proposed model could be used to classify new crime occurrence areas that are of similar nature to other well-known occurrence areas.

Even though the newly predicted areas may not have seen recent crimes, if they are similar enough to prior ones, they could be considered to be high-risk areas to monitor closely. This is an important advantage given that in some areas people are less inclined to report crimes.[44] Moreover, our approach provides new ways of describing geographical areas. Recently, some criminologists have started to use risk terrain modeling[45] to identify geographic features that contribute to crime risk, for example, the presence of liquor stores, certain types of major stores, and bars. Our approach can identify novel risk-inducing or risk-reducing features of geographical areas. In particular, the features used in our approach are dynamic and related to human activities.

Further, as suggested in the Introduction section, this study is relevant to two related debates that will shape to a great extent the future expansion and maturation of big data for social good as an intellectual and operational field: first, on the differences and complementarities between the predictive and prescriptive uses of big data, and, second, about the potential trade-offs between short- and medium- to long-term policy interventions.

By design, predictive approaches are not meant to identify and thus address the complex processes that contribute to criminal behaviors in human societies. However, they can, as in the case of our study, shed interesting light on correlates of crime that are not out of

the reach of public policies and community-based programs (e.g., specific people dynamics and characteristics of places). Unveiling such correlates may inform subsequent academic research and policy pilots that may lead to crime reduction in the long run (e.g., crime prevention through environmental design[46]). In other words, insights from predictive models may inform prescriptive approaches (and vice versa).

In addition, short-term effects can have, cumulatively, structural impacts. In the area of conflict prevention, for example, "operational" (or direct) prevention efforts in general and "preventive diplomacy" in particular have been increasingly recognized as critical to longer-term "strategic" interventions intended to address the root causes of conflicts—economic, political, etc. One argument is that short-term, targeted interventions that avoid conflict escalations allow sociopolitical adjustment mechanisms to take place, gradually gearing societies away from oscillations around various stages of violence.[47] Community-based early warning systems of conflict have also been found to be more efficient than their previous top-down counterparts,[48] and it remains to be seen whether and how "at-risk" communities could be empowered to make the most of big data for social good to reduce crime.

Note that the case study described in the article suffers from a number of limitations due to the constraints of the datasets used. First of all, we had access only to 3 weeks of Smartsteps data collected between December 2012 and the first week of January 2013. In addition, the crime data provided was aggregated on a monthly basis. As previous studies have shown, different crime types follow different temporal patterns.[49] Furthermore, having access to crime events aggregated on a weekly, daily, or hourly basis would enable us to validate the described approach with finer time granularity, predicting crime in the next week, day, or hour. Finally, the human behavioral data used in this study is derived from the mobile network infrastructure of a mobile operator. There are many other sources of human behavioral data that could also be included in our analysis (e.g., geotagged social media and public transport logs) and that could add complementary and valuable information for the task at hand. We leave this exploration to future work.

However, despite these limitations, the proposed approach illustrates the value of large-scale human dynamics data—which is actually available in an existing product (Smartsteps)—to classify crime levels.

Author Disclosure Statement

No competing financial interests exist.

References

1. Lazer D, Pentland A, Adamic L, et al. Computational social science. Science. 2009;323:721–723.
2. Letouzé E. Big data for development: Facts and figures. SciDev.Net. Spotlight: Big data for development. Available online at www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html (last accessed August 8, 2015).
3. Wesolowski A, Eagle N, Tatem A, et al. Quantifying the impact of human mobility on malaria. Science. 2012;338:267–270.
4. Frias-Martinez E, Williamson G, Frias-Martinez V. An agent-based model of epidemic spread using human mobility and social network information. In: Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011, pp. 57–64.
5. Eagle N, Macy M, Claxton R. Network diversity and economic development. Science. 2010;328:1029–1031.
6. Song X, Zhang Q, Sekimoto Y, et al. Prediction of human emergency behavior and their mobility following large-scale disaster. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2014, pp. 5–14.
7. Pastor-Escuredo D, Torres Fernandez Y, Bauer JM, et al. Flooding through the lens of mobile phone activity. Proceedings of IEEE GHTC, 2014.
8. Shang J, Zheng Y, Tong W, et al. Inferring gas consumption and pollution emission of vehicles throughout a city. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD), 2014, pp. 1027–1036.
9. Letouzé E. Big Data for Development: Challenges and Opportunities. United Nations Global Pulse: New York, NY, 2012.
10. Letouzé E, Meier P, Vinck P. Big data for conflict prevention: New oil and old fires. In: New Technology and the Prevention of Violence and Conflict. Mancini F, ed. International Peace Institute: New York, NY, 2013, pp. 4–27.
11. Morozov E. To save everything, click here: the folly of technological solutionism. PublicAffairs, 2013.
12. Toole JL, Eagle N, Plotkin JB. Spatiotemporal correlations in criminal offense records. ACM Trans Intell Syst Technol. 2011;2:1–18.
13. Wang T, Rudin C, Wagner D, et al. Learning to detect patterns of crime. In: Machine Learning and Knowledge Discovery in Databases. Springer: Berlin, 2013, pp. 515–530.
14. Ferrara E, De Meo P, Catanese S, Fiumara G. Detecting criminal organizations in mobile phone networks. Expert Syst Appl. 2014;41:5733–5750.
15. Traunmueller M, Quattrone G, Capra L. Mining mobile phone data to investigate urban crime theories at scale. In: Proceedings of 6th International Conference on Social Informatics (SocInfo). Lecture Notes in Computer Science, Springer, vol. 8851, 2014, pp. 396–411.
16. Mehlum H, Moene K, Torvik R. Crime induced poverty traps. J Dev Econ. 2005;77:325–340.
17. Cullen J, Levitt S. Crime, urban flight, and the consequences for the cities. Rev Econ Stat. 2009;81:159–169.
18. Ehrlich I. On the relation between education and crime. In: Education, Income and Human Behavior. Juster FT, ed. McGraw-Hill: New York, NY, 1975, pp. 313–338.
19. Braithwaite J. Crime, Shame and Reintegration. Cambridge University Press: Cambridge, UK, 1989.
20. Patterson EB. Poverty, income inequality, and community crime rates. Criminology. 1991;29:755–776.
21. Jacobs J. The Death and the Life of Great American Cities. Random House: New York, NY, 1961.
22. Newman P. Defensible Space: Crime Prevention Through Urban Design. Macmillian Pub Co.: New York, NY, 1972.
23. Felson M, Clarke R. Opportunity Makes the Thief: Practical Theory of Crime Prevention. Home Office, 1998.
24. Brantingham PL, Brantingham PJ. A theoretical model of crime hot spot generation. Stud Crime Crime Prev 1999;8:7–26.
25. Weisburd D. Place-based policing. Ideas Am Policing. 2008;9:1–16.
26. Boggs SL. Urban crime patterns. Am Sociol Rev. 1965;30:899–908.
27. Farrington DP, Loeber R. Some benefits of dichotomization in psychiatric and criminological research. Crim Behav Ment Health. 2000;10:100–122.

28. Streiner DL. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. Can J Psychiatry. 2002;47:262–266.
29. Shannon C. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423.
30. Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. Science. 2010;327:1018–1021.
31. Krumme C, Llorente A, Cebrian M, et al. The predictability of consumer visitation patterns. Scientific Reports. 2013;3, article 1645.
32. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;2:1157–1182.
33. Singh SR, Murthy HA, Gonsalves TA. Feature selection for text classification based on gini coefficient of inequality. J Mach Learn Res. 2010;10:76–85.
34. Tuv E, Borisov A, Runger G, et al. Feature selection with ensembles, artificial variables and redundancy elimination. J Mach Learn Res. 2009;10:1341–1366.
35. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–140.
36. Biau G. Analysis of a random forests model. J Mach Learn Res. 2012;13:1063–1095.
37. Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine learning (ICML), 2008, pp. 96–103.
38. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Lear Tech. 2011;2:37–63.
39. Provost F, Fawcett T. Data science for business: what do you need to know about data mining and data analytic thinking. O'Reilly 2013.
40. Eck J, Chainey S, Cameron J, et al. Mapping Crime: Understanding Hotspots. National Institute of Justice, 2005.
41. Mohler G, Short M, Brantingham P, et al. Self-exciting point process modelling of crime. J Am Stat Assoc. 2011;106:100–108.
42. Turvey B. Criminal Profiling: An Introduction to Behavioral Evidence Analysis. Academic Press: San Diego, CA, 1999.
43. Ellis L, Beaver KM, Wright J. Handbook of Crime Correlates. Academic Press: San Diego, CA, 2009.
44. Tarling R, Morris K. Reporting crime to the police. Br J Criminol. 2010;50:474–479.
45. Caplan J, Kennedy L. Risk Terrain Modeling Manual: Theoretical Framework and Technical Steps of Spatial Risk Assessment for Crime Analysis. Rutgers Center on Public Security: Washington, DC, 2010.
46. Jeffery CR. Crime Prevention Through Environmental Design. Sage Publications: Beverly Hills, CA, 1977.
47. Bock JG. The Technology of Nonviolence: Social Media and Violence Prevention. MIT Press: New York, NY, 2012.
48. OECD. Preventing Violence, War and State Collapse: The Future of Conflict Early Warning and Response. OECD: New York, NY, 2009.
49. Felson M, Poulsen E. Simple indicators of crime by time of day. Int J Forecasting. 2003;19:595–601.

Cite this article as: Bogomolov A, Lepri B, Staiano J, Letouzé E, Oliver N, Pianesi F, Pentland A (2015) Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. *Big Data* 3:3, 148–158, DOI: 10.1089/big.2014.0054.

Abbreviations Used

AUC = area under the ROC
GCSE = General Certificate of Secondary Education
LSOAs = lower layer super output areas
RF = Random Forests
SVMs = support vector machines