Paper Title: Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data

Paper Authors: Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nutria Oliver, Fabio Pianesi, Alex Pentland

Link to Paper: https://arxiv.org/pdf/1409.2983.pdf

Abstract: In this paper, we present a novel approach to predict crime in a geographic space from multiple data sources, in particular mobile phone and demographic data. The main contribution of the proposed approach lies in using aggregated and anonymised human behavioural data derived from mobile network activity to tackle the crime prediction problem. While previous research efforts have used either background historical knowledge or offenders' profiling, our findings support the hypothesis that aggregated human behavioural data captured from the mobile network infrastructure, in combination with basic demographic information, can be used to predict crime. In our experimental results with real crime data from London we obtain an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not. Moreover, we provide a discussion of the implications of our findings for data-driven crime analysis.

Summary: In this paper, the authors proposed a method to predict crime in a geographic area using human behavioural data derived from a combination of mobile network activity and demographic information. Until the above method was presented, most existing research work had been from a people-centric perspective and made use of prior occurrences of crimes to identify patterns of crimes committed by the same offender/group of offenders, etc. A place-centric approach for crime hotspot detection and prediction as presented by the authors complements already existing methods and contributes to criminal studies and data-driven criminal studies. The datasets used are Criminal Cases Dataset (includes geo-location of all reported crimes with month and year tags, specific location of crime and type of crime (e.g. burglary, shoplifting, violent crime, etc.)), Smartsteps Dataset (includes footfall per geographic division cell in London called Smartsteps cells and details about gender, age and home/work/ visitor group splits) and London Borough Profiles Dataset (contains 68 different metrics about the population of a geographical area(e.g. ethnicity, language, employment, etc.)). The problem is treated as a classification task to predict if a particular cell will be a crime hotspot in the next month. Since the Smartsteps cell IDs, crime locations and borough profiles are not spatially linked, each crime event was mapped to a Smartsteps cell which it occurred closest to. The same was done for the borough profile dataset. The mean, median, standard deviation, min and max values and Shannon entropy is calculated for each Smartsteps variable. The criminal dataset was split into low crime (class 0) and high crime (class 1) classes depending on whether the number of crimes in a cell was higher than the median value. 80% of data is used for training and 20% for testing the classifier. Using Pearson correlation analysis, a large subset of the features are found to have strong mutual correlations. Hence, some of the features are dropped. The mean decrease in the Gini coefficient of inequality is used to rank features and to decide which features to select. From a set of over 6000 features, a subset of 68 features that are expected to have maximum influence in minimising error. Random Forest decision tree algorithm is found to have the best performance among other methods (regression, SVMs, neural networks). Accuracy, F1 and AUC score are the performance metrics used to evaluate the classification algorithm. It is seen that higher-level features extracted over a sequence of days have more predictive power than monthly extracted data. Features extracted from people who are 'at home' are found to be of high importance. The model is able to predict a crime hotpot in the following month with an accuracy of 70%. This information can be used by the police departments to determine where to implement higher security.