Computational Criminology

Vincent Etter
LCA, EPFL
Lausanne, Switzerland
vincent.etter@epfl.ch

***Abstract*—In this report, we present our work done in spring 2011 on the UK crimes dataset. This dataset was first released in December 2010, and contains reports of crimes committed in England and Wales, with their type and location. We first perform some exploratory analysis on this data, by looking at the correlation of crime rates with some independent variables, such as the population density or the unemployment rate, as well as the relationship between different types of crimes. We also study the spatial autocorrelation of the crime rates. Then, we define a classification problem in which we are interested in identifying probable criminals from mobility traces and aggregated crimes reports. We first introduce a basic algorithm to try and solve this problem, and then reformulate our model to fit a probabilistic group testing setup.**

I. Introduction

As part of its Open Government initiative, the United Kingdom started releasing in December 2010 detailed crime records as a freely available dataset. Several applications and websites have emerged from this, allowing people to compare neighborhoods in terms of crime rates, find the safest place to move in a new town, or know which intersections to avoid when traveling in a foreign city. Information extracted from this kind of dataset is also of high interest for both city administrations and politics, as it could for example allow to organize more efficiently police forces [3].

Crime data mining is by far not a new idea. Many studies have been published on the subject, either exploring spatial correlations [12], relationship with other covariates [2], or both [7]. Various areas such as US cities, Belgian countryside or the Swedish capital have been studied, but the completeness of the UK dataset allows for the first time to perform a detailed crime data analysis at the scale of a whole country.

In section II, we first describe the structure of the dataset, and the specific fields we used in our analysis.

We study in section III the relationship between district crime counts and different covariates, such as unemployment rate or population density.

Interactions between crimes of different types are described in section IV.

In section V, we try to fit a probability distribution to the monthly crime rates of districts, and assess the goodness-of-fit of the resulting distribution.

Then, we explore spatial distribution and autocorrelation of crimes in the different UK districts in section VI.

We describe an interesting classification problem in section VII, in which we are interested in finding probable criminals in a set of people, using their mobility traces and crime counts in different areas. Two approaches are proposed to solve it, a very simple one (that obtains limited results), and a formulation as a probabilistic group testing setup.

Finally, we conclude this report in section VIII and discuss future work to be done on this dataset.

TABLE I
Total number of crimes reported in the UK, per type, from Dec 2010 to Apr 2011.

	Dec 2010	Jan 2011	Feb 2011	Mar 2011	Apr 2011
Other	144542	169056	170766	190800	184725
Vehicle	29283	34673	33424	35002	33305
Violent	57207	59302	56224	61844	63596
Burglary	37825	44328	41789	43739	40005
Robbery	5679	6592	6399	6495	6225
Anti-social	201520	202536	207600	241942	277341
Total	476056	516487	516202	579822	605197

II. Dataset

Since December 2010, the UK government releases monthly a dataset containing all crimes that were reported during the previous month. This data is freely available on their website[1], and contains several pieces of information about each crime, among which a type and a location. Six different types of crimes are reported: violent crimes, vehicle crimes, robberies, burglaries, anti-social behaviors and other crimes. The location of a crime is given as the street on or near which it was committed[2]. Moreover, spatial coordinates (corresponding usually to an arbitrary point along the street) are given, allowing to easily plot crimes on a map, and to compute distances between them.

Table I shows the total number of crimes reported in the UK for each month and each type of crime. While these numbers may seem high, one should note that anti-social behaviors and other crimes account for the majority of reports, and these represent mostly benign crimes, such as hoax calls and shoplifting. Figure 1 shows a boxplot of the proportion of crimes of each type reported monthly in the UK, from Dec 2010 to Apr 2011.

We see that while the number of crimes fluctuates slightly (even when taking into account the number of days per month), the proportion of crimes of each type is fairly stable.

[1] http://www.police.uk/data

[2] This introduces some uncertainty, in order to protect the victim's privacy.

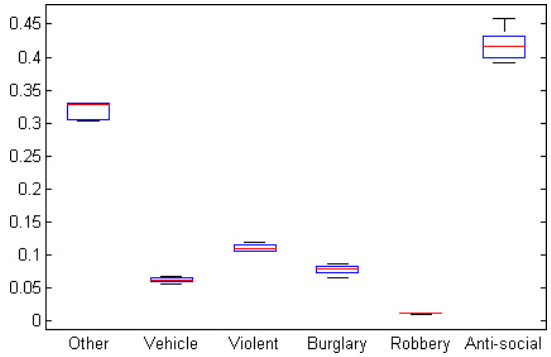


Fig. 1. Boxplot of the proportion of crimes of each type reported each month in the UK, between Dec 2010 and Apr 2011.

A. District counts

While the fine spatial resolution of the data is useful for studying spatial distribution, other covariates are usually not available at such a precise level. Indeed, information like the mean yearly income or the unemployment rate is rather published for bigger entities, for instance cities, districts or counties. Thus, we had to resample our data in order to generate a dataset at a coarser level, by counting the number of crimes of each type using some tessellation of the UK territory.

In order to have the finest resolution as possible, we decided to use administrative districts as spatial units. These are the smallest regions for which most statistics are available. There are 380 of them in the UK. We used the official districts boundaries[3] to assign each crime to a district, given its spatial coordinates.

B. Crime counts versus crime rate

Up to this point, we only considered crime counts as our variable of interest. However, a better statistic to use is the crime rate, that is the number of crimes per habitant (or per thousands of habitants). Indeed, using the number of crimes directly to compare districts is not a good idea, as districts more populated are more likely to have a higher number of crimes that districts with a smaller population. Thus, we will exclusively use crime rates instead of raw crime counts when comparing districts.

Let us define the crime rate for crime type t in a district d during month m as:

$$C_{t,d,m} = \frac{X_{t,d,m}}{N_d}$$

where $X_{t,d,m}$ is the number of crimes of type t reported in district d during month m, and N_d is the latest estimated population of district d. In other words, $C_{t,d,m}$ is the average number of crimes of type t per habitant of district d during month m.

We can then define the mean crime rate of district d for type t as:

[3]Available at http://www.ordnancesurvey.co.uk/oswebsite/products/os-vectormap-district/index.html

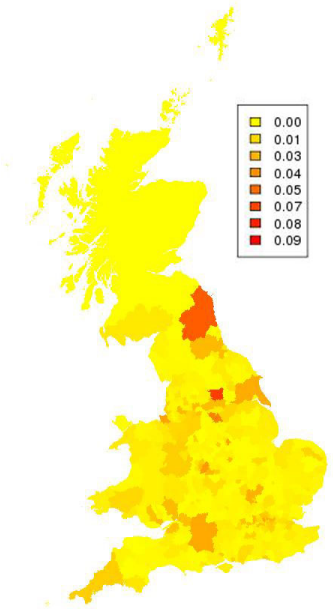


Fig. 2. Mean total crime rates for each district, defined for one district as the total number of crimes reported in this district during the whole time period, divided by the district population times the number of months.

TABLE II
SUMMARY OF THE DISTRICTS' STATISTICS USED AS INDEPENDENT VARIABLES.

	Min.	Median	Mean	Max.
Mean yearly income (£)	20850	29740	31970	92230
Unemployment rate (%)	2	7.1	7.23	15.4
Proportion of white (%)	50.5	98.38	95.26	99.74
Proportion aged 16-24 (%)	6.2	13.9	14.09	24.4
Population density (hab/km^2)	8.016	507.3	1381	13720

$$C_{t,d} = \frac{1}{M}\sum_{m=1}^{M} C_{t,d,m}$$

where M is the number of months of data available.

To illustrate this, figure 2 shows the mean total crime rate $C_d = \sum_{t=1}^{T} C_{t,d}$ for each districts d.

III. COVARIATES

Once we obtained the crime rates of different types for each UK district, we were able to study their correlation with other covariates. More precisely, we chose the following independent variables: mean yearly income, unemployment rate, proportion of white ethnicity, proportion of population aged 16-24 and population density (as an indicator of rurality). All these covariates were taken from the latest data available on the UK official labour market statistics website[4]. Table II shows a summary of these covariates.

Following [9], we defined the log crime rate:

$$R_{t,d} = log(C_{t,d} * 1000 + 1)$$

[4]https://www.nomisweb.co.uk/

TABLE III

PEARSON'S CORRELATION AND CORRESPONDING P-VALUES OF $R_{t,d} = log(C_{t,d} * 1000 + 1)$ WITH DIFFERENT COVARIATES, WHERE $C_{t,d}$ IS THE MEAN NUMBER OF CRIMES OF TYPE t PER HABITANT OF DISTRICT d. BOLD VALUES MEAN A CORRELATION HIGHER THAN 0.3, AND RED P-VALUES INDICATE INSIGNIFICANT RESULTS ($p \geq 0.05$).

	Mean income	Unemp. rate	Prop. white	Prop. 16-24	Pop. density
Other	0.11 / 0.05	0.11 / 0.04	-0.20 / 0.00	0.16 / 0.00	0.29 / 0.00
Vehicle	0.18 / 0.00	0.22 / 0.00	**-0.38** / 0.00	0.17 / 0.00	**0.40** / 0.00
Violent	0.05 / 0.36	0.17 / 0.00	-0.27 / 0.00	0.16 / 0.00	**0.34** / 0.00
Burglary	0.10 / 0.07	0.16 / 0.00	-0.26 / 0.00	0.16 / 0.00	0.28 / 0.00
Robbery	**0.33** / 0.00	**0.39** / 0.00	**-0.71** / 0.00	0.22 / 0.00	**0.73** / 0.00
Anti-social	-0.07 / 0.22	0.13 / 0.02	-0.11 / 0.05	0.14 / 0.01	0.18 / 0.00

and studied its correlation with the different covariates described above. To do so, we used Pearson's correlation, defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

For n samples of paired data (X_i, Y_i), it can be written as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $\bar{X}$ is the sample mean of X.

Table III shows the resulting correlations, with the corresponding p-values. We see that the number of robberies correlates with most of the covariates, and that urban areas have higher rates of vehicle crimes, violent crimes and robberies.

We should also note the high negative correlation of the proportion of white ethnicity in the population with robberies and vehicle crimes. However, we will not venture an explanation about this result.

IV. RELATIONSHIP BETWEEN CRIME TYPES

We also investigated the relationship between different types of crime, to verify if most of the crimes happen together, regardless of their type, or if some types occur more than others in some districts. Figure 3 shows a matrix plot of $R_{t,d}$, for different pairs of crime types $(t_i, t_j), i \neq j$. Scatter plots allow to visually assess linear dependencies between sample pairs. Clearly, except for robberies, we see that all types of crime appear to have some direct linear relationship. This relationship is illustrated by the cloud of points falling close to a diagonal line in the plots.

This fact is confirmed by the correlation $corr(R_{t_i,d}, R_{t_j,d})$ for $i \neq j$, shown in table IV. Except for robberies, that are slightly less correlated to other crime types (but still strongly correlated with $r \geq 0.75$), all other types are highly correlated.

A. SVD and dimensionality reduction

To investigate further this separation of the different types, we applied a dimensionality reduction technique. We modeled each district as a point in a T-dimensional space, where each dimension corresponds to a type of crime. Then, we defined

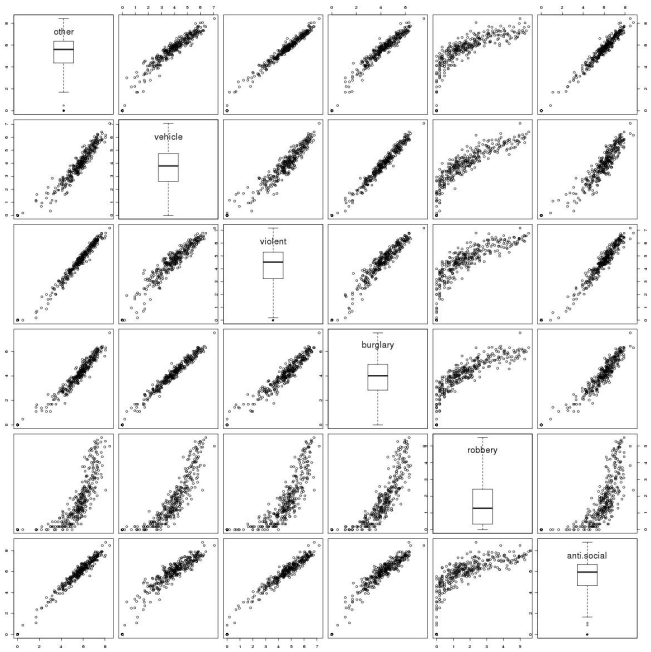


Fig. 3. Matrix plot of $R_{t,d}$ for different pairs of crimes types $(t_i, t_j), i \neq j$. Types in order are: other, vehicle, violent, burglary, robbery and anti-social. This illustrates graphically the relationship between different types of crimes. Each line and column has respectively the same type for its y and x axis. Diagonals show scatter plots of $R_{t,d}$.

TABLE IV

$corr(R_{t_i}, R_{t_j,d})$ FOR $i \neq j$, *i.e.* CORRELATION BETWEEN THE LOG CRIME RATES OF DIFFERENT TYPES, ALONG WITH ASSOCIATED P-VALUES.

	Other	Vehicle	Violent	Burglary	Robbery
Vehicle	0.97 / 0.00				
Violent	0.99 / 0.00	0.97 / 0.00			
Burglary	0.98 / 0.00	0.99 / 0.00	0.98 / 0.00		
Robbery	0.75 / 0.00	0.85 / 0.00	0.79 / 0.00	0.82 / 0.00	
Anti-social	0.99 / 0.00	0.95 / 0.00	0.98 / 0.00	0.97 / 0.00	0.71 / 0.00

a $D \times T$ data matrix $\mathbf{D}$ where each line represents a district, each column a type of crime, and a cell $\mathbf{D}_{i,j} = \sum_{m=1}^{M} X_{j,i,m}$ counts the total number of crimes of type j reported in district i.

We applied a SVD decomposition to this matrix $\mathbf{D}$. The resulting singular vectors, along with their corresponding singular values, are shown in table V.

We see that the first component is more or less proportional in all dimensions, meaning that it captures the scale of the data, but no individual relation. The second singular vector,

TABLE V

SINGULAR VECTORS AND ASSOCIATED SINGULAR VALUES OF THE SVD DECOMPOSITION OF $\mathbf{D}$, WHERE $\mathbf{D}_{i,j}$ IS THE TOTAL NUMBER OF CRIMES OF TYPE j REPORTED IN DISTRICT j.

Other	0.425	-0.177	-0.182	-0.520	0.655	-0.234
Vehicle	0.428	0.072	0.379	0.308	-0.172	-0.737
Violent	0.428	-0.044	-0.212	-0.502	-0.714	0.092
Burglary	0.418	-0.138	0.686	0.028	0.108	0.569
Robbery	0.344	0.865	-0.227	0.152	0.139	0.198
Anti-social	0.400	-0.441	-0.506	0.599	0.019	0.176
Sing. value	2.28	0.70	0.39	0.27	0.19	0.14

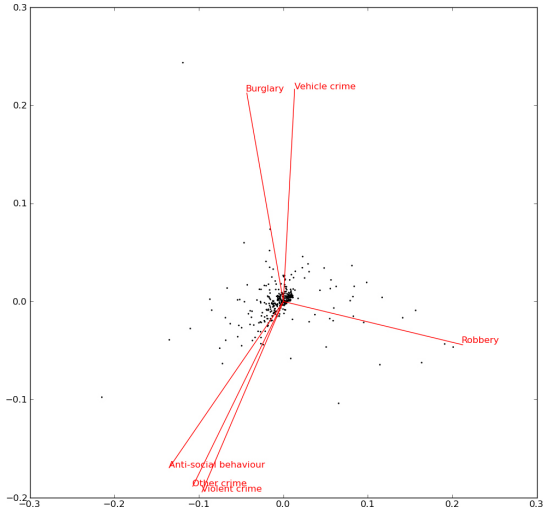


Fig. 4. Projection of the matrix **D** on its second and third singular vectors. The axis of the original 6-dimensional space, corresponding each to one type of crime, are shown in red, illustrating the separation between some of the types.

however, separates clearly robberies from other types, having a much greater component that the others. While this already appeared during the correlation analysis, the reason behind robberies standing out is not clear to us. Intuitively, one could say that a robbery is a crime much more involved than breaking a window to enter an empty house, or shoplifting. Thus, it must be committed by a small fraction of the population, that lays low the rest of the time and commits less other crimes.

Similarly, the third component puts vehicle crimes and burglaries together and opposed to anti-social behavior. We won't push our social analysis further, except to note these relationships.

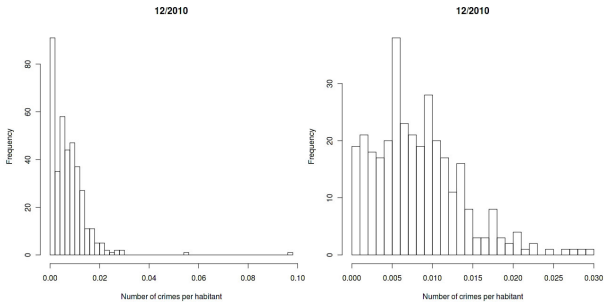
To illustrate these separations, we projected the data on the second and third singular vectors, as a mean of dimensionality reduction [10]. Figure 4 shows the resulting figure. The projection of the axis of the original 6-dimensional space clearly show the separation of the different types mentioned above.

V. Distribution fitting

Finally, we were interested in trying to fit a probability distribution to the monthly crime rate in each district $C_{d,m} = \sum_{t=1}^{T} C_{t,d,m}$. Being able to model precisely this crime rate would have many useful applications, for instance to generate new data for simulations.

We proceeded in a systematic way, and started by looking at the shape of the histograms of the crime rates distribution. Figure 5a shows such an histogram, for crimes reported in Dec 2010. We noticed that there is a large concentration of districts having a crime rate of zero, which corresponds to missing values. Thus, we removed these entries, as well as two outliers that have very high crime rates, and obtained the histogram showed in figure 5b.

According to the histogram shape, we focused on distributions of the exponential family, and tried to fit several of them to our data. Our best candidate was the Weibull distribution



(a) Original (b) Cleaned

Fig. 5. Distribution of the monthly crime rates $C_{d,m}$ for all districts, in Dec 2010. (a) shows the original histogram, (b) shows the histogram after having removed zeros (missing values) and two outliers.

[11]. This distribution is characterized by two parameters, the shape k and slope λ, and is defined as follows:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

Using maximum-likelihood methods, we fit this distribution to the crime rates $C_{d,m}$ of each month. The values we obtained for these parameters were almost exactly the same for each month: $k = 1.45, \lambda = 0.01$. Figure 6 shows the corresponding Q-Q plot of the monthly crime rate $C_{d,m}$ for Dec 2010 and Mar 2011 as examples (other months have similar Q-Q plots).

We verified the goodness of fit of the distribution with these parameters using a Kolmogorov Smirnov test. This test quantifies a distance between the empirical distribution function of the samples and the cumulative distribution function of the reference distribution, in this case a Weibull.

For each month, the null hypothesis was accepted with $p > 0.10$, meaning that the Weibull distribution cannot be rejected as the true distribution. Other goodness-of-fit tests exist, such as Anderson-Darling or χ^2, and should be applied to further assert the correctness of our choice of distribution and parameters.

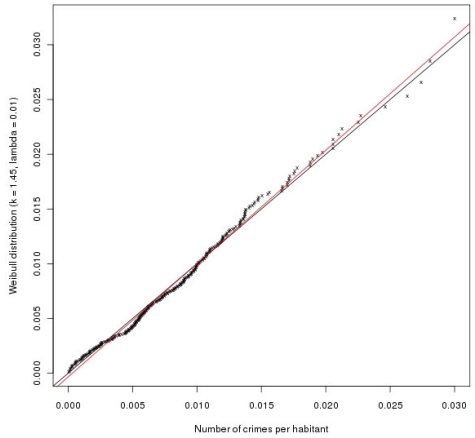
VI. Spatial autocorrelation

All the relationships we studied up to this point focused on districts as entities, but did not take at all into account the spatial dimension. Indeed, it would be interesting to see if a higher crime rate in a given district has some influence on its neighbors. Or, in other words, do criminals travel, or are they inspired by what others close to them do?

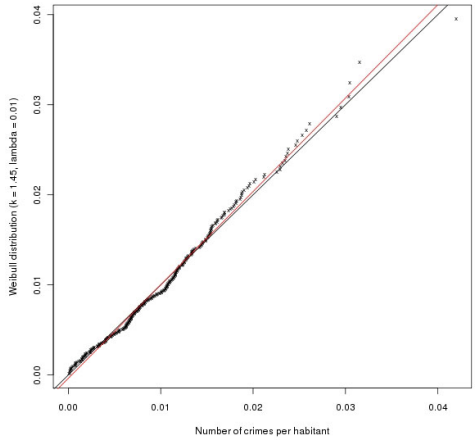
To try and answer this question, we explored the global spatial autocorrelation of the mean crime rates $C_{t,d}$. To do so, we used two different measures: Moran's I [8] and Geary's C [6]. Let us first define the $D \times D$ matrix of spatial weights $\mathbf{W}$. It captures the spatial relations between districts as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if districts } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

Moran's I is defined as follows:



(a) Dec 2010



(b) Mar 2011

Fig. 6. Q-Q plot of the monthly crime rates in all districts $C_{d,m}$ versus a Weibull distribution, for Dec 2010 and Mar 2011. The parameters of the distribution were obtained using a maximum likelihood method: $k = 1.45, \lambda = 0.01$ in both cases. The red line is the $x = y$ diagonal, the black line is the linear regression of the data points.

$$I = \frac{N}{\sum_i \sum_j \mathbf{W}_{ij}} \frac{\sum_i \sum_j \mathbf{W}_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where X is the variable of interest, $\bar{X}$ the sample mean, N the number of samples and $\mathbf{W}_{ij}$ the weights matrix.

Moran's I ranges from -1 (indicating perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern.

Geary's C is defined as:

$$C = \frac{(N-1)\sum_i \sum_j \mathbf{W}_{ij}(X_i - X_j)^2}{2W \sum_i (X_i - \bar{X})^2}$$

where W is the sum of all $\mathbf{W}_{ij}$.

It is inversely related to Moran's I, but not identical, as it is more sensitive to local spatial autocorrelation. Its values lie between 0 and 2, where 1 means no spatial autocorrelation, 0 a positive and 2 a negative one.

A. Results

Table VI shows the resulting measures on the mean crime rates $C_{t,d}$ across the different UK districts, for each type of crime. We see that for all but one type, both I and C are significantly close to the "no autocorrelation" value, meaning that there is no spatial relationship between the districts. Robberies stand out one more time with a slightly higher spatial autocorrelation, suggesting that authors of such crimes may travel to nearby districts to commit other offenses.

TABLE VI
MEASURES OF GLOBAL SPATIAL AUTOCORRELATION (AND ASSOCIATED P-VALUE) OF THE MEAN CRIME RATES $C_{t,d}$ ACROSS ALL UK DISTRICTS, FOR DIFFERENT CRIME TYPES, USING MORAN'S I AND GEARY'S C. VALUES CLOSE TO RESPECTIVELY 0 AND 1 INDICATE NO SPATIAL AUTOCORRELATION.

	Other	Vehicle	Violent
Moran's I	0.12 / 0.00	0.20 / 0.00	0.14 / 0.00
Geary's C	0.95 / 0.19	0.87 / 0.00	0.88 / 0.00
	Burglary	Robbery	Anti-social
Moran's I	0.13 / 0.00	0.60 / 0.00	0.06 / 0.05
Geary's C	0.96 / 0.38	0.43 / 0.00	1.06 / 0.13

These two measures assume a certain homogeneity amongst districts, meaning that either they all are correlated, or none are. As our results above prove neither of these cases, this homogeneity assumption may not hold, and thus we may see some districts correlated in parts of the country, and some completely independent in other parts.

To test this, there exists a local indicator of spatial association (LISA) [1] that computes a local Moran's I for each spatial element and allows to evaluate its statistical significance. It is defined as:

$$I_i = \frac{X_i}{m_2} \sum_j \mathbf{W}_{ij} X_j$$

where $m_2 = \frac{1}{N}\sum_i X_i^2$.

These local estimators showed interesting results, as seen in figure 7, in which we plotted the I_i values for districts with significant p-values ($p < 0.05$).

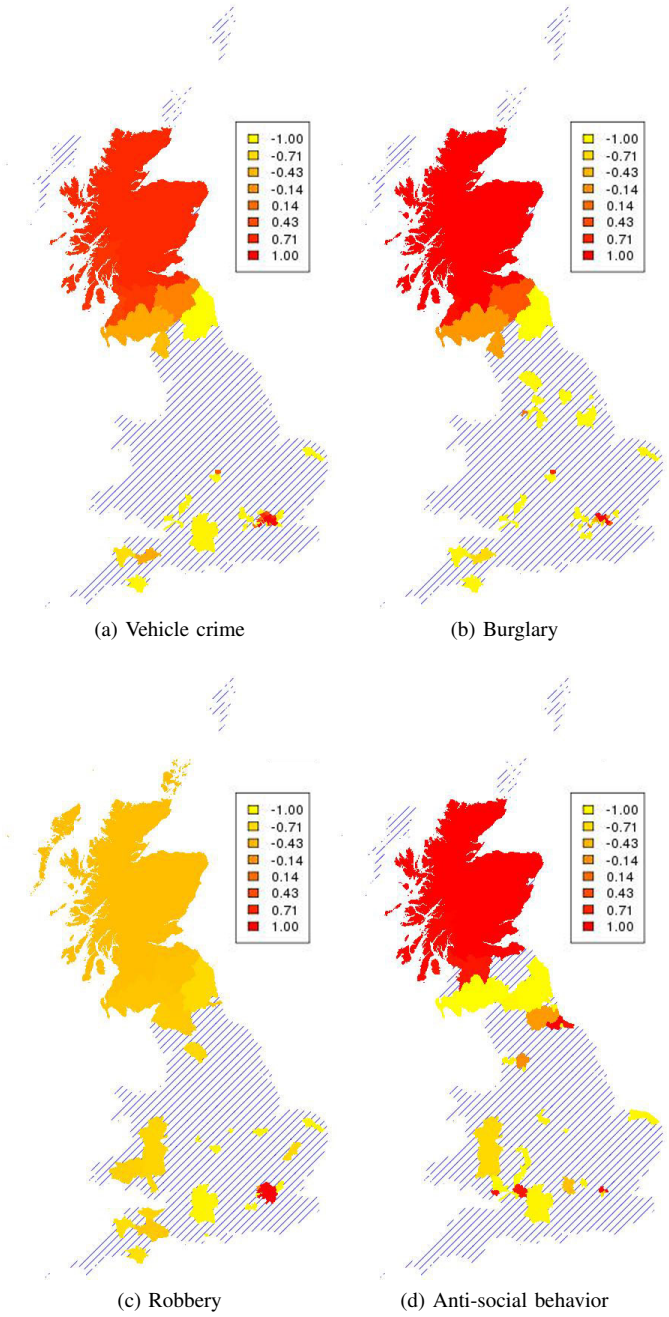
First, note that the upper districts show a strong local autocorrelation, for the reason that their crime rate is mostly zero, due to a low population.

More important are the clusters that appear on the main land, especially around London. We clearly see for instance that robberies are highly spatially correlated around the capital, suggesting that robbers from London tend to stay in town or around the suburbs, but do not wander far in the countryside.

We can see similar "clusters" for other types of crimes, notably one around Middlesbrough for anti-social behaviors. This makes sense, as Middlesbrough had the 4th highest crime rate of England in 2007.

VII. CLASSIFICATION

Once the exploratory analysis of the data was completed, we wanted to go past the mere description of a dataset, and formulate a more challenging problem. Having such precise locations of the crimes (to the street level), we wondered if it would be feasible to detect probable criminals in a set of people, knowing where each person has been during a given time period and where crimes were reported.



(a) Vehicle crime (b) Burglary

(c) Robbery (d) Anti-social behavior

Fig. 7. Values of the local indicator of spatial association (local Moran's I) for the mean crime rate $C_{t,d}$ for different crime types. High values (red) indicate a positive correlation, low values (yellow) a negative one, and intermediate values (orange) no spatial correlation. Districts striped in blue have insignificant indicators ($p > 0.05$).

This problem relies of course on the assumption that we have a way of knowing the whereabouts of people. Fortunately, this kind of mobility traces is increasingly available, either thanks to the multitude of location-aware smartphones that people carry all day long, or simply through the logs of mobile operators. However, this data is not always very precise. For example, triangulation from GSM towers usually achieves a maximal precision of fifty meters in urban areas.

Moreover, as said in section II, crime locations in this dataset are voluntarily "blurred" to protect the privacy of people, resulting in aggregates of crimes at some places, and more generally in some imprecision of the location. Thus, one should rather use a coarse resolution when dealing with this classification problem, instead of relying on finer but imprecise locations.

A. Model

To tackle this classification problem, we first started with some simulated data. We defined our space as a grid of $G \times G$ cells. In this grid, we place uniformly at random N people. Each person is a criminal with probability p, *i.e.* $P_i \sim Ber(p)$. These are sampled at the beginning of the simulation, and are then fixed.

Each person walks across the grid following a slightly modified 2D random walk: at each step, a person chooses a cell uniformly in a $(2S+1) \times (2S+1)$ square around its current position (thus allowing people to move at different speeds around the grid).

In other words, the position (x_i, y_i) of person i at time step $k+1$ is defined as:

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + U(-S, S) \\ y_i^{(k+1)} &= y_i^{(k)} + U(-S, S) \end{aligned}$$

where $U(a, b)$ is a discrete uniform random variable over the interval $[a, b]$.

Finally, at each time step $k \in (1, \ldots, T)$, each criminal i has a probability q of committing a crime:

$$C_i^{(k)} = \begin{cases} \sim Ber(q) & \text{if } P_i = 1 \\ 0 & \text{if } P_i = 0 \end{cases}$$

B. Simulation input/output

To summarize, our simulation takes the following inputs:

- G: dimension of the grid
- S: step size of each person at each time step
- T: number of time steps to simulate
- N: number of people to simulate
- p: probability of each person to be a criminal
- q: probability of each criminal to commit a crime at each time step

It outputs the following data:

- $people$: vector of N elements where $people(i) = 1$ if person i is a criminal, and $people(i) = 0$ otherwise
- $crimes$: $G \times G$ matrix giving the number of crimes committed in each cell
- $trajectories$: $T \times N \times 2$ matrix giving the position of each person i at each time k, *i.e.* $trajectories(k, i, \cdot) = \left(x_i^{(k)}, y_i^{(k)}\right)$

Figure 8 shows an example of the output of such a simulation, for $G = 50$, $S = 3$, $T = 50$, $N = 20$, $p = 0.2$ and $q = 0.2$.

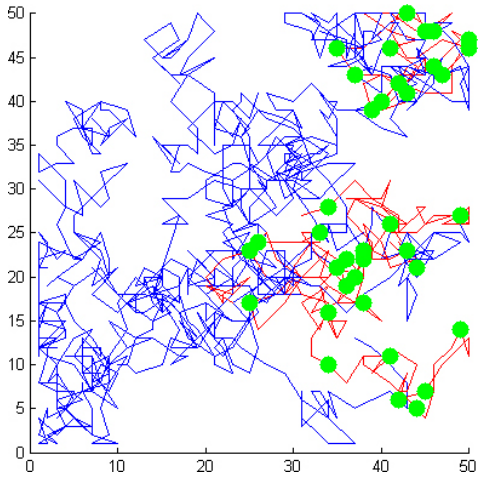


Fig. 8. Simulation of mobility traces and crimes for $N = 20$ people in a 50×50 grid during $T = 50$ time steps, with step size $S = 3$, proportion of criminals $p = 0.2$ and crime rate $q = 0.2$. Blue lines represent traces of normal people, red lines traces of criminals, and green dots crimes.

C. Recovery algorithm

From this setup, we derived a simple recovery algorithm, that aims at finding the most likely suspects from the output described above. The basic idea is the following: people that were (often) in cells where crimes occurred are more likely to be criminals.

To explain our algorithm, we will use a simple case. We consider a 2×3 grid, with $N = 3$ people, one of which is a criminal. They moved around the grid for $T = 4$ time steps, during which three crimes were committed. Figure 9 illustrates the resulting setup.

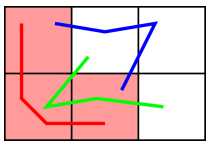


Fig. 9. Simple setup for the explanation of our recovery algorithm. $N = 3$ people walked for $T = 4$ time steps around a 2×3 grid. The red, blue and green lines show respectively the traces of each person, and the cells with red background show cells where a crime was committed.

From the traces, we can build the following $N \times H$ presence matrix $\mathbf{M}$, in which each row corresponds to a person, and each column a cell of the grid (H is here the total number of cells in the grid) :

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} 1 & 0 & 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \\ \mathbf{C}' &= \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{aligned}$$

In other words, $\mathbf{M}$ tells how many times each person was in each cell. Above is also shown the $H \times 1$ matrix $\mathbf{C}$, that counts the number of crimes committed in each cell.

Then, we simply normalize each column of $\mathbf{M}^{(1)}$ such that it sums to one:

$$\begin{aligned} \mathbf{M}^{(1)} &= \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{3} & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & \frac{1}{3} & 0 \end{pmatrix} \\ \mathbf{C}' &= \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{aligned}$$

Finally, we compute the score of each person by multiplying $\mathbf{M}^{(1)}$ with $\mathbf{C}$:

$$\mathbf{P}_{scores} = \mathbf{M}^{(1)} \cdot \mathbf{C} = \begin{pmatrix} 1.5 \\ 0.67 \\ 0.83 \end{pmatrix}$$

This score can be seen as the likelihood of a person being a criminal. Note that by multiplying with $\mathbf{C}$, we make sure that cells with no crimes have no influence.

Identifying the criminals is now simply a matter of taking the highest K scores, where K is the number of criminals (in our toy example, $K = 1$). Hence, in this case the first person is the criminal:

$$\mathbf{P}_{est} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

If K is unknown, we can either threshold this score, or estimate K using the proportion of criminals p: $\mathbb{E}(K) = Np$.

D. Results

We tested the efficiency of our algorithm with the following setup: $G = 100$, $N = 1000$ people, $T = 100$ time steps, step size $S = 5$, probability of being a criminal $p \in [0, 1]$ and crime rate $q \in [0, 1]$.

We measured the performance as follows:

$$E(\mathbf{P}_{est}, \mathbf{P}_{real}) = hd(\mathbf{P}_{est}, \mathbf{P}_{real})/N$$

where $hd(\mathbf{x}, \mathbf{y})$ is the hamming distance between $\mathbf{x}$ and $\mathbf{y}$.

This metric counts the proportion of false classifications, *i.e.* both false positives (regular people wrongly categorized as criminals) and false negatives (criminals not detected).

We measured this error for different values of p and q, and plotted the results in figure 10.

These results show that for high crime rates, our algorithm performs quite well. However, for low crime rates, it is unable to identify properly criminals, simply because most of them committed very few or even no crimes at all. This is clearly illustrated when the crime rate is zero: even while no crimes were committed, our algorithm still tries to identify criminals, and thus ends up assigning people arbitrarily.

To verify if it scaled properly, we fixed the proportion of criminals and the crime rate, and compared our algorithm with a random guess, for an increasing number of people N. Figure 11 shows the results.

We see that with small number of people, our algorithm performs significantly better than a random guess. However, the error seems to increase with the number of people. This is most probably due to the combinatorial explosion of the possible assignments. Indeed, the grid being only 100×100,

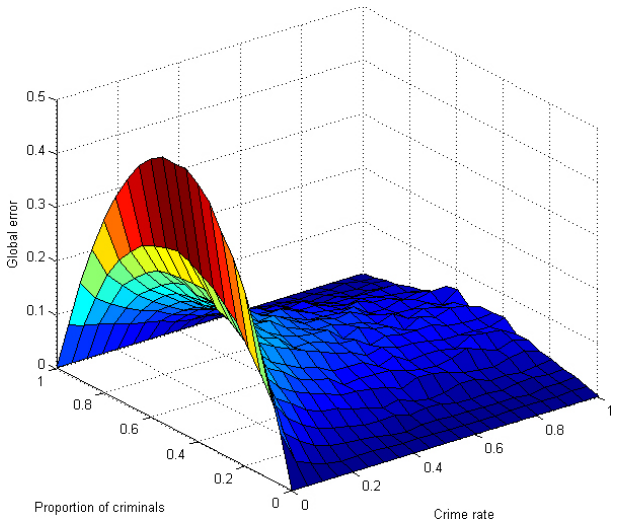


Fig. 10. Performance of our algorithm on simulated data on a 100×100 grid, with $N = 1000$ people, during $T = 100$ time steps, with a step size $S = 5$. 21 values of both the criminals proportion p and crime rate q were tested, and the proportion of wrong classifications is plotted.

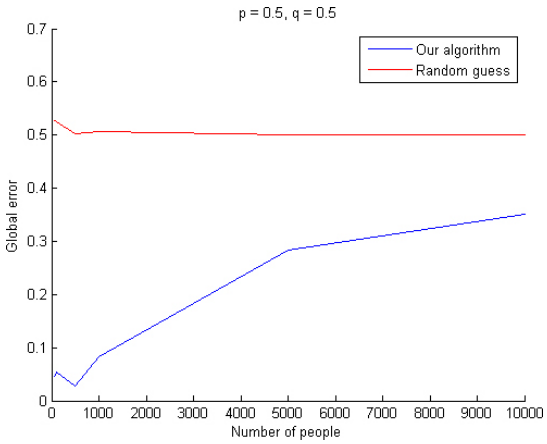


Fig. 11. Comparison of our algorithm with a random guess on a 100×100 grid with a step size $S = 5$, during $T = 100$ time steps, for $N = [50, 100, 500, 1000, 5000, 10000]$ people. The proportion of criminals is $p = 0.5$ and the crime rate is $q = 0.5$. The blue line is our algorithm, the red line is the random guess.

a high number of people simulated during $T = 100$ time steps will cover most of the grid, meaning that there will have been several people in each cell of the grid. Thus, when computing the scores, many people will have similar scores, and the algorithm will have no mean of splitting the ties.

Another observation about this algorithm is that it uses only "local" knowledge to make its decisions, in the sense that cells have no influence on the others. Indeed, if one cell where a crime was reported only had one person visiting it, this would mean that this person is undoubtedly a criminal. Hence, a more efficient algorithm should take advantage of this fact during the reconstruction process, which ours does not.

E. Group testing

Inspired by [4], we tried a slightly more complex approach to model our reconstruction problem, by using a probabilistic group testing setup. Group testing is targeted at reducing the cost of finding certain elements of a set. It was first introduced in 1943 [5], when the US army was looking for a way to detect syphilis in groups, without having to test each soldier individually.

In classical group testing, the setup is the following: we have N items, in which at most $K \ll N$ are defective. We then test M subsets of the items, and get a test result $\mathbf{y}$:

$$\mathbf{y}_i = \begin{cases} 1 & \text{if subset } i \text{ contains } \geq 1 \text{ defective items} \\ 0 & \text{otherwise} \end{cases}$$

The subsets are defined by the following contact matrix $\mathbf{M}^{(c)}$:

$$\mathbf{M}^{(c)}_{i,j} = \begin{cases} 1 & \text{if test } i \text{ includes item } j \\ 0 & \text{otherwise} \end{cases}$$

This contact matrix simply tells which items were in which test pool. Then, group testing simply aims at finding a sparse vector $\mathbf{x}$ such that:

$$\mathbf{y} = \mathbf{M}^{(c)} \cdot \mathbf{x}$$

The novelty introduced by [4] is to add some uncertainty to the testing procedure: each defective item has a probability q of being detected. This is the same as generating a sampling matrix $\mathbf{M}^{(s)}$ by randomly flipping the ones of $\mathbf{M}^{(c)}$ to zero with probability $1 - q$.

In other words, each element of $\mathbf{M}^{(c)}$ is mapped to the corresponding element of the sampling matrix $\mathbf{M}^{(s)}$ by passing through the channel shown in figure 12.

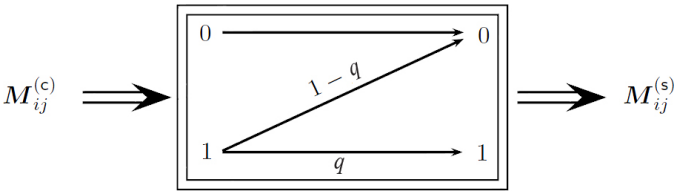


Fig. 12. Each element of the sampling matrix $\mathbf{M}^{(s)}$ is generated independently from the corresponding element of the contact matrix $\mathbf{M}^{(c)}$ by passing through a channel that flips ones to zeros with probability $1 - q$.

F. Application to our problem

We feel that this setup could be adapted to our initial classification problem:

- each cell corresponds to a test
- the pool of item on which the test is performed corresponds to the people that were present in the cell
- the test outcome is the number of crimes that were committed in the cell during the T time steps
- each element of $\mathbf{M}^{(c)}$ would be flipped to zero with a probability $p(1 - q) + (1 - p)$, which is the probability that a person does not commit a crime.

We should check if the required conditions presented in [4] are satisfied by our setup, and then implement their algorithm. Unfortunately, this remains to be done.

VIII. Conclusion

During this project, we explored a new dataset that was made available in December 2010, and that has continued to grow ever since. It contains monthly reports of crimes committed in the UK, with their type and location.

First, we described the principal characteristics of this dataset, and defined crime rates in districts as our variable of interest.

We studied its correlation with different covariates, and found that the urbanization (represented by the population density) was an important factor affecting the number of vehicle crimes, violent crimes and robberies. We also noticed that the proportion of white people correlates negatively with the rate of vehicle crimes and robberies, but did not push the reasoning further. Finally, we noted that robberies correlated with nearly all covariates, making this type of crime stand out from others. This finding was corroborated by the analysis of the correlation of the different crime types, which revealed that all types but robberies are highly correlated with each other, meaning that there is not a type that tends to happen on its own.

This separation of types appeared again when we applied a dimensionality reduction technique to the dataset. Indeed, it split the types in three groups, with robbery on its own, burglary and vehicle crime as a pair, and the three remaining types (violent crime, anti-social behavior and other crime) together. We put forward an explanation for this separation, by differentiating these three groups of crimes with respect to their severity (a robbery is a crime much more involved than a graffiti) and the amount of contact between the criminal and the victim (burglaries and vehicle thefts usually happen when the victim is absent).

We were able to fit a Weibull distribution to the monthly crime rates in each district, allowing us to have a good model of this variable. However, we do not have an intuitive or formal justification of the reason why these crime rates follow this distribution.

Then, we explored the spatial autocorrelation of the crime rates between the districts. Using global measures of autocorrelation, we did not uncover any homogeneous pattern. We were nevertheless able to identify a few cluster of districts that have significant spatial correlation between them for some crime type, either positive or negative. This was notably the case for robberies, for which districts in the London area were highly autocorrelated.

Finally, we proposed an interesting problem of classification, in which we would like to identify the most likely suspects when given mobility traces of users and report of crimes. We proposed a simple algorithm to solve this problem. It obtained fair results, but scaled poorly with the number of people, merely due to the combinatorial explosion of possible classifications. Instead, we introduced the notion of probabilistic group testing, and explained how such an approach could be applied to our problem.

A. Future work

Now that the principal characteristics of this dataset are more familiar, it is a good starting point for more advanced enquiries:

- **Temporal prediction**: for now, only five months of data are available. While this is plenty for spatial analysis, having only a monthly temporal resolution is pretty limiting. Indeed, it would be interesting to try and apply recurrent neural networks such as an echo state network (ESN) or a liquid state machine (LSM) to perform time-series prediction. The problem with such learning approaches is that they usually require large numbers of training samples, which translates to several months of data.
- **Spatial analysis**: we obtained some interesting results of local autocorrelation of the crime rates, but these results were restricted to districts, which still represent large areas. It would be interesting to apply the same analysis at a much smaller resolution. This would allow to answer interesting questions such as "Can one go from a safe neighborhood to a nest of criminals by simply crossing a street?", or "Are different types of crimes restricted to some neighborhoods?".
- **Criminals detection**: our algorithm presented in section VII-C was very simple, but still performed fairly well in some conditions. There is room for a lot of improvement, for instance in taking the crime rate into account, or by using information about known criminals to compute the likelihood of other suspects.
- **Group testing**: in section VII-F, we proposed a way of applying probabilistic group testing to our criminals identification problem. However, we formulated it, but did not verify if the necessary conditions of the contact matrix were fulfilled, nor did we implement the reconstruction algorithm.

References

[1] Luc Anselin. Local indicators of spatial associationlisa. *Geographical Analysis*, 27(2):93–115, 1995.

[2] Mitchell B. Chamlin and John K. Cochran. An excursus on the population size-crime relationship. *Western Criminology Review*, 5:119–130, 2004.

[3] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, and M. Chau. Crime data mining: a general framework and some examples. *Computer*, 37(4):50 – 56, april 2004.

[4] Mahdi Cheraghchi, Ali Hormati, Amin Karbasi, and Martin Vetterli. Group Testing with Probabilistic Tests: Theory, Design and Application. *IEEE Transactions on Information Theory*, 2010.

[5] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[6] R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):pp. 115–127+129–146, 1954.

[7] Marc Hooghe, Bram Vanhoutte, Wim Hardyns, and Tuba Bircan. Unemployment, inequality, poverty and crime. *British Journal of Criminology*, 51:1–20, 2011.

[8] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):pp. 17–23, 1950.

[9] D. Wayne Osgood. Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16:21–43, 2000. 10.1023/A:1007521427059.

[10] Michael E. Wall, Andreas Rechtsteiner, and Luis M. Rocha. Singular Value Decomposition and Principal Component Analysis. *ArXiv Physics e-prints*, pages 91–109, August 2002.

[11] Waloddi Weibull. A statistical distribution function of wide applicability. *ASME Journal of Applied Mechanics*, 18:293–297, 1951.

[12] Haifeng Zhang and Michael P. Peterson. A spatial analysis of neighborhood crime in ohama, nebraska using alternative measures of crime rates. *Internet Journal of Criminology*, 2007.