115th European Study Group with Industry

Barcelona, January 2016

Edited by

Tim Myers, CRM
Joan Solà-Morales, UPC
Maria Aguareles, UdG
Marta Pellicer, UdG

Centre de Recerca Matemàtica
Campus de Bellaterra, Edifici C
08193 Bellaterra
Barcelona, Spain

First edition: Juliol 2017

ISSN (printed edition)
ISSN (electronic edition)

Legal deposit:

Preface

The 115th European Study Group with Industry took place at the Centre de Recerca Matemática between 25-29 January, 2016.

The problems presented at the meeting were chosen to provide a wide variety of subject areas and to appeal to local academics. Specifically, the four problems concerned were:

- The manufacture of nanocrystals.
- Predicting burglaries in Catalunya.
- Equilibria in Tokamak plasmas.
- Waiting time for inventory.

Approximately 50 people attended the meeting, coming from as far afield as France and Girona. There were also participants from Ireland, Bulgaria and Georgia. Funding was primarily through the COST Action, TD1409, Mathematics for Industry Network (MI-NET) as well as the Math-In Network, the Barcelona Graduate School of Mathematics, the Catalan Maths Society and the Càtedra Lluís A. Santaló of the University of Girona.

Tim Myers, Barcelona 2016

Organising Committee

Tim Myers, Centre de Recerca Matemàtica
Joan Solà-Morales, Universitat Politècnica de Catalunya
Maria Aguareles, Universitat de Girona
Marta Pellicer, Universitat de Girona
Ernest Benedito, Universitat Politècnica de Catalunya
Àlex Haro, Universitat de Barcelona
Carlos Vázquez-Cendón, Universidade da Coruña

Contents

Searching for a Predictive Model for Burglaries in Catalonia

Problem presented by

Pere Boqué[1]

Report prepared by

Maria Aguareles[2,9], Dusan Bikov[3], Marta Canadell[4], Enric Costa-Miracle[5], Carlos Domingo[6], Jesús Fernández-Sánchez[7], Núria Folguera[5], Jordi Saludes[7], Aleksandra Stojanova[3], Zlatko Varbanov[8]

Study group contributors

Maria Aguareles[2,9], Dusan Bikov[3], Marta Canadell[4], Enric Costa-Miracle[5], Carlos Domingo[6], Jesús Fernández-Sánchez[7], Núria Folguera[5], Jordi Saludes[7], Isabel Serra[5], Aleksandra Stojanova[3], Zlatko Varbanov[8]

Abstract: In this report we present a summary of the discussions and results obtained during the 115th European Study Group with Industry on the prediction of burglaries in Catalonia. This problem was presented by the Police Department and the goal was to obtain models for the dynamics of burglaries and to infer spatial and temporal patterns based on data from 2010 until 2015.

1 Statement of the problem

Mossos d'Esquadra is the name of the Police Department in Catalonia who have competences for prevention and investigation of criminal actions concerning civil security. In the last years burglaries have been one of the most relevant offences that they have dealt with. These are serious offences against citizens privacy which, in many occasions, end up with

[1]Mossos d'Esquadra de Catalunya, Sabadell, Spain
[2]Universitat de Girona, Girona, Spain
[3]Goce Delcev University, Stip, Macedonia
[4]Brown University, USA
[5]Centre de Recerca Matemàtica, Barcelona, Spain
[6]Universitat de Barcelona, Barcelona, Spain
[7]Universitat Politècnica de Catalunya, Barcelona, Spain
[8]University of Veliko Turnovo, Bulgary
[9]Report co-ordinator: maria.aguareles@udg.edu

intimidation and violence. Furthermore, these type of crimes take place all around the catalan territory and in all sort of environments, they occur in both flats and houses and at any time of the year. Recently, the Mossos d'Esquadra have been especially interested in the possibility of preventing burglaries by using prediction models.

Some cities around the world (see for instance [6]) have already implemented civil security strategies based on mathematical predicting tools. However, these tools strongly rely on a certain degree of regularity in the city planning and in the fact that offenders tend to choose their next target mainly according to a geographical proximity. This is, in fact, the major drawback when one tries to implement these models in Catalonia, where the territory is very heterogeneous. Cities in Catalonia combine very different types of homes, and it is very common to find neighbourhoods with detached houses and dense sets of flat buildings. Also, the structure and urbanisation of large cities and small towns is very different. According to police experience, burglars tend to specialize in particular types of households and so they move according to parameters which go beyond the physical proximity of houses. For instance, burglars often choose to burglarize places with fast escaping routes which are close to highways, or houses with particular types of fences or doors, etc.

In this report we present the results of the analysis of burglary data provided by *Mossos d'Esquadra* and we propose different approaches to derive predicting models, thus improving the algorithms that the police are now using. We start by exploring if geographical proximity patterns (the so-called *near-repeat victimisation*) can be deduced from the data. We then first consider small regions, of the level of small cities, to devise if waves of burglaries take place in concentrated periods of time. We thus remove the spatial dependence of the data in this first approach. As we shall show, several features arise from this first analysis: first, we observe that patterns arise only if a certain level of layer information is used, and secondly we observe that if one considers small towns or cities, it is indeed possible to predict the lapse of time between two consecutive waves of burglaries in the town. We then broaden our analysis to the possibility of having non-trivial spatial patterns of criminal activity. In particular we propose an algorithm to derive connections between different regions of Catalonia, that is to say, the goal is to detect if two given sets of burglaries which have taken place in regions that may be far away from each other have any relation at all, in the sense that when a burglary takes place in, for instance, region A, it is highly probable that burglaries in B will also take place. Finally we propose a model

2 Results

2.1 Exploring repeat-victimisation

The algorithms presented by the Mossos d'Esquadra in their initial exposition were based on the theory of repeat (or near-repeat) victimization. This criminology theory states that it is expectable that a target (in our case, a household or an area) suffers from repeated criminal victimization in a short period of time. *Near repeats* refer to targets with similar characteristics or situations, in the sense that the metric of the model is not necessarily determined only by spatial distance. This idea can be modeled by means of the so-called Hawkes process. The Hawkes process is a self-exciting point process that was introduced in seismology to model earthquakes and their aftershocks. More precisely, one assumes that there is a background rate at which events are likely to occur, and once an event does occur, the rate jumps up and aftershocks are expected to follow. See [2] for an application of this theory to gang rivalries in Los Angeles. The original algorithm that the Mossos d'Esquadra presented divided the territory in rectangular cells. Then, using data from the previous two weeks, they determined the potential risk of burglary in each cell for the following days. The algorithm heavily relied on the assumption that there was a near-repeat victimization phenomenon. However, the consideration of uncategorized events can hide these patterns. For instance, if we represent the burglaries in Sant Feliu de Guíxols[2] over time, it is hard to see any temporal cluster supporting the idea of repeated victimization (see Figure 1).

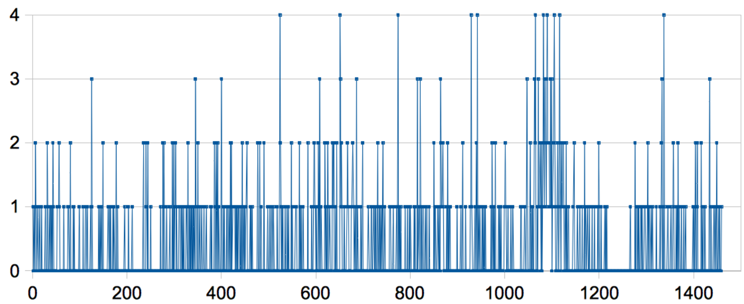

Figure 1: Total number of burglaries in Sant Feliu de Guíxols. 2011–14.

Given the information that we had at hand about each burglary (mainly, the type of construction), we decided to plot the same diagram but only for events that took place in country houses (see Figure 2). In this case, one can observe that there are temporal clusters that may be explained by a near-repeat victimization model, even in the case of Girona[3] city (see Figure 3).

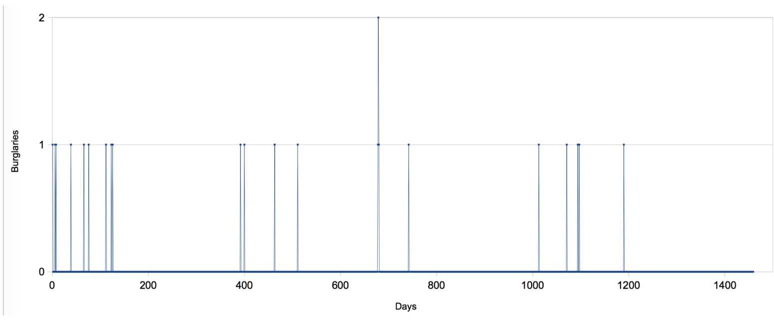

Figure 2: Burglaries in Country Houses in Sant Feliu de Guíxols. 2011–14.

Sticking to the same category of country houses, we also plotted the burglaries in the whole region of Girona[4]. In this case, the presence of temporal clusters is not as obvious (see Figure 4), so even if we are considering a particular type of burglary, the size of the region is still important if we want to find repetition patterns.

[2] 16,23 km^2 and 21.810 inhabitants.
[3] 39,14 km^2 and 97.227 inhabitants
[4] 5.905 km^2 and 756.810 inhabitants.

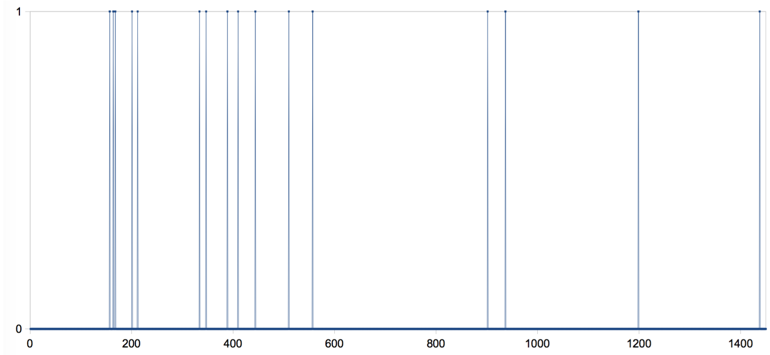

Figure 3: Burglaries in Country Houses in Girona City. 2011–14.

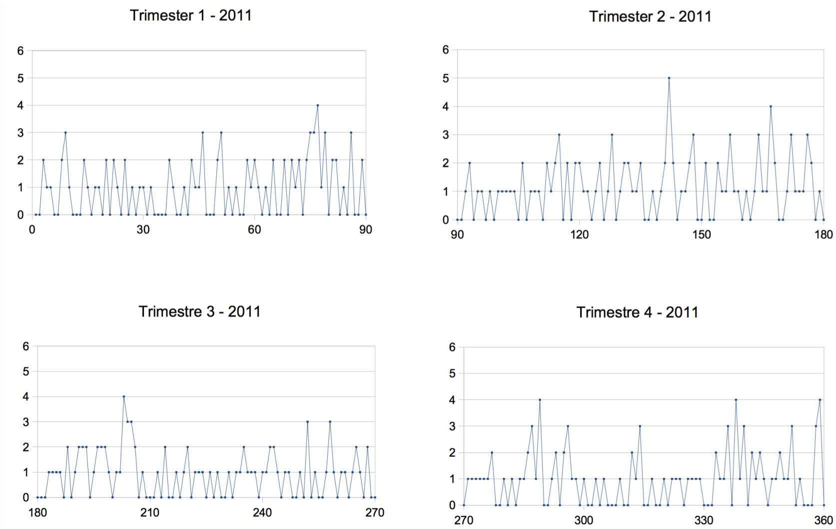

Figure 4: Burglaries in Country Houses in the Girona region. 2011 in trimesters.

A similar test can be carried out with other types of households (apartments, or houses), but these categories seem to be too broad (see Figure 5).

However, we are confident that if we could add additional data such as loot type in order to create subcategories, there is a chance that we could observe patters as in the country houses of Sant Feliu de Guíxols. Summing up, it could be useful to

- Define spatial cells adapted to the territory, using qualitative information such as the orography or demography of the area.
- Categorize burglaries in a few classes that can help us classify different kinds of events.
- Define different layers in each cell consisting of households (or areas) that could be potential targets of each class of burglary.

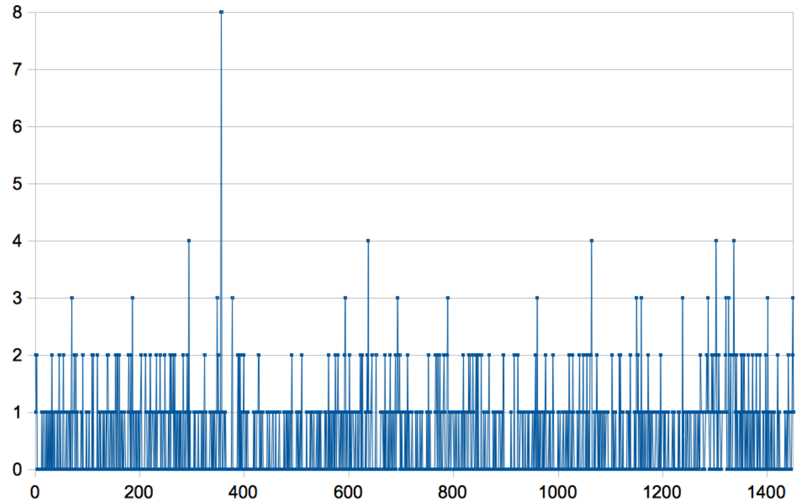

Figure 5: Burglaries in Apartments in Girona City. 2011–14.

- Run a near-repeat victimization based algorithm on each layer of each cell independently, using just the data of the corresponding class.

2.2 Time-dependent statistical analysis

To efficiently manage police resources, the Mossos d'Esquadra need some information of when and where the next wave of burglaries is most likely to take place. As a first approximation we have clustered the information at the level of cities and towns and we have computed the elapsed time between consecutive waves of events. The goal is to devise for each town whether there is a mean for this waiting time between events, which would indicate how long after a wave of burglaries in a city or town a new wave is expected.

We have used data corresponding to years from 2010 to 2014 and we have analysed it by means of the coefficient of variation, CV. The coefficient of variation, CV, is a standardized measure of dispersion of a probability distribution, that is to say, it explains the extent of variability in relation to the mean of the data studied. More precisely, it is defined as

$$CV = \frac{\sigma}{\mu}, \tag{1}$$

where σ is the standard deviation and μ is the mean of the sample.

In general, when studying this ratio, three main cases must be distinguished depending on how far from unity this quantity is:

- $CV \ll 1$ implies that the variance is very small with respect to the mean and therefore the events take place in almost a periodic way. For instance, this could imply that a burglaries are occurring every two days or once per week (the same day of the week). In this case, the corresponding density function is a Dirac delta and there clearly is a time pattern in the burglaries.
- $CV \simeq 1$ implies that the mean and the standard deviation are of the same order. In this case, when the standard deviation of an exponential function is equal to

its mean, the waiting time distribution function decays like an exponential. Exponential density functions correspond to Poisson processes which have been used to model near-repeat victimisation models (see for instance [3]). This models assume that if there is an offence in a given area, the probability that the burglars may strike in the same place or nearby is higher than in other places. This high probability prevails for a short time and it rapidly falls afterwards.

- When $CV \gg 1$ the data has a high-variance with respect to the mean. In this case, the distribution function for the waiting times does not decay exponentially but it has a fat tail. This is due to the presence of cluster formations in time since the events are self-attracting or self-activating. This would correspond to an scenario where the waves of burglaries last longer than just one day and they are followed by a longer period of peace.

We then want to compute the values for the coefficient of variation for some cities and towns in Catalonia. We first start by removing the zero waiting time events from the data, or, in other words, we don't take into account if a given day one or more events have taken place. The coefficient of variation after having removed the 0-waiting time events is usually known as the coefficient of variation of the residuals (CVR).

When computing the CVR for all the cities/town/villages available in the data, we observe the three different regimes explained above. For instance, the city of Sabadell, which is a medium size city, is found to have a CVR well bellow one. Amposta, which is a smaller town, has $CVR \simeq 1$ and El Bruc, that is even smaller, has $CVR >> 1$. In general we observe that large cities tend to have $CVR \ll 1$ but its mean is 1 day, which simply says that there are burglaries every day. This, in fact, does not give any information in terms of prevention and it is clear that it is due to the high population density in these cities. It is clear that in these cases one must consider smaller areas, probably at the level of neighbourhoods. At the other end, very small towns seem to have CVR well over one. It makes sense that in these towns or villages clusters appear since these locations are not so well-connected by public transport or main roads. Then, once the burglars move there, they stay around for a period of time. In table 1 we show the names of the cities in Catalonia where the CVR has been found to be greater than 1 and the ones where it has been found to be bellow 1:

$CVR \ll 1$	$CVR \ll 1$	$CVR \gg 1$
Barcelona	Sabadell	Begur
Tarragona	Reus	Cabrils
Girona	Santa Coloma de Gramenet	Creixell
Lleida	Mataró	El Bruc
Terrassa	Badalona	Sant Hilari Sacalm

Table 1: Cities with a $CVR \ll 1$ (1st and 2nd colum) and with $CVR \gg 1$ (3rd column).

Having a quick look at the table and knowing a little bit about Catalonia, the first feature that can be deduced is that large cities like Barcelona tend to have a very defined mean ($CVR \ll 1$), but, as in the case of Barcelona in all these cities the mean is of one day. In order to check this feeling in a more analytical way, we decided to search at the Catalan Institute of Statistics (Idescat) the number of inhabitants of these cities. In Table 2, we can see the number of inhabitants of the 11th largest cities in Catalonia:

An approach that the Mossos d'Esquadra were using was, in fact, to divide the whole Catalonia in cells and to investigate the evolution of events in each of these cells according to a near-repeat victimisation approach. In their simulations they already noticed that a crucial point is the size of such cells. In this sense, they were interested in understanding the optimal size of a cell in terms of predictions. Our conclusion with this results is that the same cell size cannot be used in all Catalonia and large cities require a more local analysis. As an example we have also computed the CVR for each neighbourhood in Barcelona and, in this case the CVR becomes of order one, thus implying that a near-repeat victimisation might be taking place. There are also other cities, like Granollers, where Mossos d'Esquadra had explicitly checked that near-repeat victimisation models where giving good predictions. Our conjecture is that only in areas where the CVR is close to one such near-victimisation models will be useful. In this sense, the CVR may be used as a tool to decide which cities require a more local analysis.

Position in the ranking	City	Number of inhab.
1	Barcelona	1.604.555
2	L'Hospitalet de Llobregat	252.171
3	Badalona	215.654
4	Terrassa	215.214
5	Sabadell	207.814
6	Lleida	138.542
7	Tarragona	131.255
8	Mataró	124.867
9	Santa Coloma de Gramenet	116.950
10	Reus	103.194
11	Girona	97.586

Table 2: Largest cities in Catalonia, IDESCAT, 2015.

To sum up we can conclude several things. The first one is that there seems to be a clear relationship between the number of inhabitants of the place we are studying and the value of the CVR. Large cities have $CVR << 1$ and thus, the near and repeat victimization pattern cannot be applied straightforward since it will not work. A new methodology should be designed and it seems that splitting the city into neighbourhoods and then applying the cells' model could be useful.

For medium size cities, that is cities that have between 10000 and 100000 inhabitants, which normally have $CVR \simeq 1$, near-repeat victimization models should suit well as it happened with Granollers. For the towns and villages, usually with $CVR >> 1$, we know that there are going to be clusters but we cannot assume that a near-repeat victimisation model will work. In this cases there might be connections with cells which are which are easily reached through highways, for instance. This idea will be more detailed in the following section.

2.3 Exploring spatio-temporal patterns

Repeat victimisation does not only occur in the shape of hot-spot. That is to say, there might be burglars who prefer to operate in the vicinity of highways and so they strike in places that are kilometres apart from each other. We want to infer these type of relations by exploring the data. To this aim, we have tried to design a model as simple as possible that allows to predict burglaries in terms of the data provided by the mossos. We propose the following algorithm:

- Start by partitioning the region in N cells (the smaller, the most accurate).
- We now define $f(i, j, \ell)$ as the "influence" of cell j on cell i during a time span of ℓ (for instance, ℓ could be 1 or 2 weeks):

$$\begin{aligned} f(i, j; \ell) := &\text{ number of events in cell } j \text{ occurring in a period of} \\ &\ell \text{ time steps after an event occurs in } i \end{aligned}$$

- A contingency table (or matrix) $F(\ell)$ is constructed in this way for each value of ℓ. The bigger the entries the more related the cells i and j are.
- By normalizing the rows of this matrix so that each sums to one, we obtain a transition matrix $P(\ell)$, whose entries are estimates of the conditional probablilities

$$\begin{aligned} p(i|j; \ell) = &\text{ probability of an event in cell } j \text{ after an event } i \text{ in a} \\ &\text{time window of length } \ell. \end{aligned}$$

- After a burglary event occurs in one cell, say i, the maxima among the entries of the corresponding row i of $P(\ell)$ should indicate the cells where it is more probable that another burglar event occurs in the following ℓ weeks. Moreover, studying the monotony of the sequence $\{p(i|j; \ell)\}_\ell$ we can measure how long is the influence of cell j in cell i.

Remark 1. A similar analysis can be performed restricting data to events recorded after some particular date k_0. The corresponding tables $F(\ell; k_0)$ and $P(\ell, k_0)$ should allow conclusions about the influence between cells in terms of the seasons or periods of the year. For example, different weeks of the year (different dates k_0) may produce different influence between cells. Parameters k_0 and ℓ allow different analysis and would play an important role for the analysis.

Remark 2. One could apply statistical analysis (namely, compute the chi-square statistic) to test if the data contained in the table $F(\ell)$ reveal dependence or independence between the events registered. Another approach would be to estimate how far the contingency table is from having rank one (this meaning independence among variables). To this aim, one can compute the singular values $\{\sigma_m\}_m$ of the matrix $F(\ell)$ and estimate the distance of $F(\ell)$ to the space of matrices of rank 1 as $\sqrt{\sum_{m\geq 2}\sigma_m^2}$. This kind of analysis should allow to validate or reject the near-repeat victimization assumption between cells or in a specified group of cells (for example, cells in the area of Barcelona).

This procedure should be improved by considering data associated to 1st or 2nd residence, or data of the metropolitan area and data of less-populated territories separately. As stated before, some amount of layer information might be helpful to detect patterns.

2.4 Graph theory

In this section we describe a method to simulate street networks in a city and to use them to detect possible hotspots. The idea is to think of the network map as a graph. The classical way to model the street network map is: the vertices of the graph represent the intersections between the streets and the links represent the streets which connect the intersections. A positive integer number (called *weight*) can be assigned to any link. In classical models the weight is the physical length of the street but for the purpose of this work this weight can be defined as a value of a function with more arguments: the physical length, the density of the flats/houses, the number of people living there, etc.

We shall denote by $G = (V, E)$ a graph (network) where $V = \{v_1, \ldots, v_n\}$ is a non-empty set of n vertices (nodes) and $E \subseteq V \times V$ is a set of links (edges) which connect them. In figure 6 we show an example of a network represented by a graph.

Finding hot spots One method to find possible hot spots for crimes (especially burglaries) was presented by T. Davies and S. Bishop (2013). A *path* in a network is defined as an ordered sequence of vertices such that there exists a link between any two consecutive vertices. The *length* of the path is the number of links (for an undirected graph) or the sum of weights of these links (for a directed graph). The *shortest path* between two vertices i and j (if there exists paths between them) is the one with minimal length. The so called *Betweenness centrality* is a measure which quantifies how often individual links are used during journeys through the network. These are the main steps to calculate this measure:

1. initialize all links with a betweennes centrality of 0;
2. consider all pairs of nodes i and j;
3. for each pair i and j, find the shortest path(s) between them;
4. for every link that appears in the shortest path(s), increment its betweenness centrality by $1/w$, where w is the number of shortest paths between i and j

More formally, if σ_{ij} is the total number of shortest paths between i and j, and $\sigma_{ij}(e)$ is the total number of shortest paths between i and j which contain the link $e \in E$, the betweenness centrality $C^b(e)$ of a given link can be defined as:

$$C^b(e) = \sum \frac{\sigma_{ij}(e)}{\sigma_{ij}}$$

for all vertices i, j such that there exists a path between i and j.

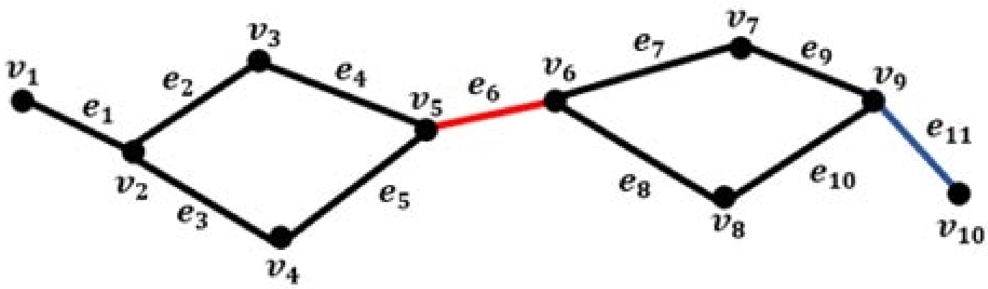

Figure 6: Example network

Spatio-temporal patterns Actually, in this case, instead of a model for a street network we need a model for the *event's (burglaries) network*. Such model was presented by T. Davis in his PhD Thesis (2015). This model can be firstly represented by means of two networks: one is based on the distance between two events (spatial proximity) and the other one is based on the time period between the events (temporal proximity). In these two networks the events are vertices but the corresponding sets of links are defined in different ways. If we take two threshold values D (spatial radius) and T (temporal radius) then

$$E_d^D = \{(i,j) \mid d_{ij} \leq D\},$$

and

$$E_t^T = \{(i,j) \mid 0 < t_{ij} \leq T\}.$$

Here, if two events occur in a given time period T, they are connected by a directed link from the earlier event to the later. This is the main difference: G_d^D is an undirected graph but G_t^T is a directed graph.

The undirected graph G_d^D and the directed graph G_t^T contain all the relevant information related to the set of events. By analysing both networks G_d^D and G_t^T, couples of events i and j which are close in space and time can be found. If the events are close in both space and time then the corresponding vertices are adjacent. One can now create an specifically event directed network G_{dt}^{DT}, of pairs which are close in space and time. This is the event network for the dataset, and it includes a space-time clustering analysis and it shows relations between events. This new network is the intersection of the spatial and temporal networks, that is,

$$E_{dt}^{DT} = \{(i,j) \mid (i,j) \in E_d^D \text{ and } (i,j) \in E_t^T\}.$$

The relationship between these three networks G_d^D, G_t^T, G_{dt}^{DT} is presented in Figure 7. Event network G_{dt}^{DT} is constructed using the other two, but it does not include the whole information for events from them. For example, if two events are not linked in G_{dt}^{DT}, it is impossible to know the reason for this disconnection, it could be because there is not a link in G_d^D, or G_t^T, or in neither of them.

In Figures 7 and 8 we provide an example with a simple set of event's data . We can see there the relation between the spatial network G_d^D, the temporal network G_t^T and the network of space-time pairs or events network G_{dt}^{DT}. The links of the event network G_{dt}^{DT} appear in the same position for both the spatial network G_d^D and the temporal network G_t^T.

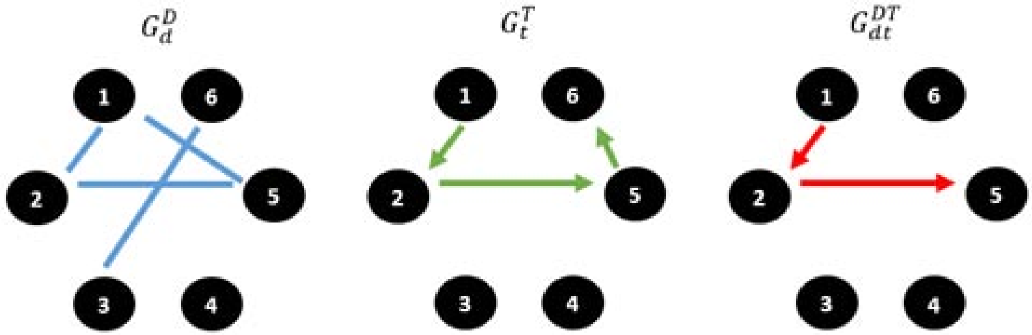

Figure 7: Original data, relationship between the three networks G_d^D, G_t^T and G_{dt}^{DT}

Figure 7 shows the connection for the original data while figure 8 shows how this can be changed with a simple permutation $\sigma(where 2 \leftrightarrow 6)$ of temporal data. This can be used in a Knox test with a Monte-Carlo approach.

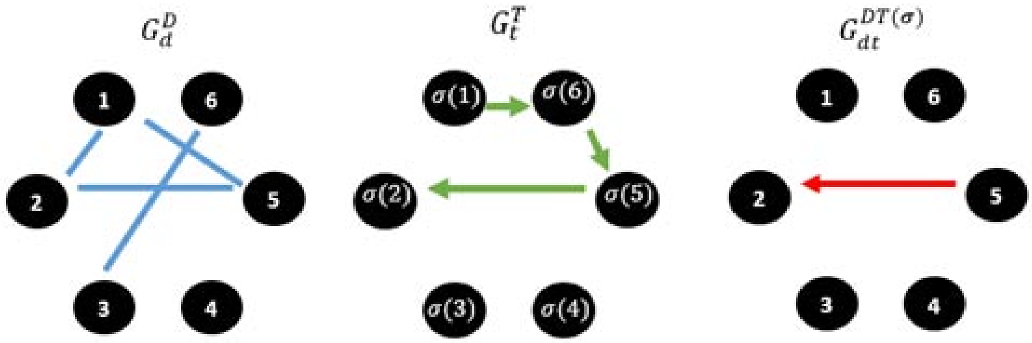

Figure 8: Relationship between the three networks under permutation σ of the temporal data

The Knox test is widely used in studies to detect space-time clustering or interactions between spatial and temporal distributions of a set of events. The origin of the test is in the study of childhood leukaemia. According to this concept, events are more likely to be close in space when they are close in time, and vice versa. The basis for the test is the concept of existing a close pair of events (close in time and close in space).

Acknowledgements

I. Serra acknowledges Generalitat de Catalunya for funding under grant 2014-SGR-1307 and the Spanish Government for funding under grant MTM2012-31118 . Her work is supported by a grant of the "la Caixa" Foundation on *Collaborative Mathematics* awarded to the Centre de Recerca Matemàtica. N. Folguera acknowledges the Spanish Ministry for Science and Innovation (MICINN) for funding under grant MTM2011-29342 and Generalitat de Catalunya for funding under grant 2014SGR1307. Her work is supported by a grant of the “la Caixa" Foundation on *Collaborative Mathematics* awarded to the Centre de Recerca Matemàtica. E. Costa-Miracle acknowledges the Spanish Ministry for Science and Innovation (MICINN) for funding under grant MTM2011-29342 and Generalitat de Catalunya for funding under grant 2014SGR1307. His work is supported by the BGSMath under grant MDM-2014-0445-04.

Bibliography

[1] J.T. Ornstein and R.A. Hammond. The Burglary Boost: A Note on Detecting Contagion Using the Knox Test. Journal of Quantitative Criminology (2106), 1–11.

[2] M. Egesdal, C. Fathauer, K. Louie, J. Neuman, G. Mohler, and E. Lewis. Statistical and stochastic modeling of gang rivalries in Los Angele. SIAM Undergraduate Research Online **3** (2010), 72–94.

[3] M.B. Short, M.R. D'Orsogna, V.B. Pasour, G.E. Tita, P.J. Brantingham, A.L. Bertozzi, and L.B. Chayes. A statistical model of criminal behavior. Math. Models Methods Appl. Sci. **18** (2008), 1249–1267.

[4] T.P. Davies and S.R. Bishop. Modelling patterns of burglary on street networks. Crime Science **2** (2013), 10.

[5] `http://nij.gov/topics/technology/maps/pages/crimestat.aspx`.

[6] `http://www.theguardian.com/cities/2014/jun/25/`.