Assessment of data mining methods for forensic case data analysis[1]

A.-L. Terrettaz-Zufferey[*], F. Ratle[**], O. Ribaux[*], P. Esseiva[*], M. Kanevski[**]

[*] University of Lausanne, School of Criminal Sciences, Batochime, 1015 Lausanne
anne-laure.terrettaz@unil.ch
olivier.ribaux@unil.ch
pierre.esseiva@unil.ch
** University of Lausanne, Institute of Geomatik and Risk Analysis, Amphipole, 1015 Lausanne
frederic.ratle@unil.ch
mikhail.kanevski@unil.ch

Abstract

The role of data mining in crime analysis is to detect all types of relevant structures in a dataset. Once filtered, those structures may point to previously unseen criminal activities. Data mining covers a wide range of methods and techniques whose potential strongly depends on the available dataset and the nature of the activity from which it is derived. Determining the most promising techniques to be applied in regards to different available forensic case databases is the aim of the research program.
A specific application to illicit drug profiling has been assessed. Analytical laboratory techniques are systematically applied to samples of cocaine seized by the police in order to extract their chemical profile. The data obtained is collated within a database. Cutting agents found are of particular interest, because they result from a treatment occurring toward the end of the distribution process. Their interpretation may then provide information on possible local illicit traffic networks. By focusing on the co-occurrence of set of cutting agents, relevant patterns are detected. This combinatorial approach using graph theory will be further tested on other crime data.

1. Introduction

Data mining is a term widely used and it is the object of many definitions. It can be understood as the extraction of previously unknown and potentially useful information or knowledge from large datasets. The main principle is to devise computer programs that scan databases and automatically seek deterministic patterns. The potential of data mining technologies strongly depends on the nature of the available dataset. They are successfully applied in different professional fields, e.g., remote sensing, biometry, speech recognition or business and marketing. They have seldom been tested on forensic case data, whereas they may potentially help to detect unknown and evolving criminal behaviours difficult to anticipate or predict through universal models.
The intrinsic difficulty related to the use of such data lies in its heterogeneity, which is essentially caused by the lack of harmonisation across jurisdictions. Moreover, factors that help evaluate the relevancy of implementing data mining techniques range from the activity from which the dataset results to its quality (degree of uncertainty, precision, completeness).
The aim of this study is to select applications of data mining methods and techniques on forensic case data, specifically on cocaine chemical profiles (focus on cutting agents), that provide relevant results in a criminal intelligence perspective.

[1] **This research is supported by the Swiss National Science Foundation (No K-12-0-115965)**

2. General process of data mining and its forensic application

The general data mining process is constituted of three main steps (Figure 1). The **data preprocessing** step includes preliminary treatments such as data cleaning, normalisation, standardisation, and coding. The **pattern recognition step** applies statistical or mathematical methods in order to detect structures hidden in the basis data and highlight information. The **interpretation step** is the use of expert knowledge and contextual information in order to associate meaning to the detected patterns. Thus, the whole methodology is based on a cooperative effort between a human expert and a computer.

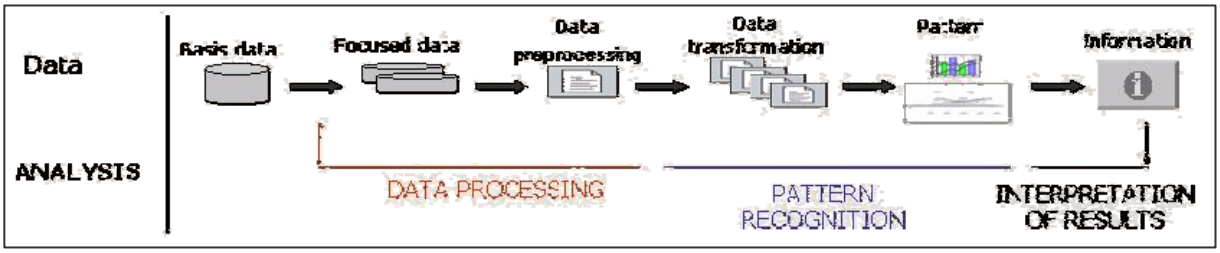

Figure 1 : General process of data mining

Law enforcement agencies and forensic science laboratories collate a large amount of data of different types resulting from criminal activities. Forensic case data is a subset consisting of the physical material collected (biological marks, toolmarks, fingermarks, shoemarks, illicit drug seizures, etc.). This kind of data can be numerical, binary (0 or 1) and categorical (for instance a color). The features extracted are frequently imprecise (essentially due to measurement tools), incomplete (fragmentary), and uncertain. In this context, the use of data mining techniques necessitates a careful pre-assessment of their potential and even an adaptation of existing pattern recognition methods (Figure 2).
This data is rarely collated within a single database. They are rather organised in relation with the type of crime under consideration: high volume crime, illicit drug trafficking, organised crime, etc. Treatments are thus to be considered along specific information processes. Additionally, the ratio signal/noise can be very weak, necessitating a careful preparation of data and extraction of a chosen set of discriminating variables. Moreover, depending on the criminal activity, structures within the database may have a very low interest. A categorisation of the detected patterns in three groups is proposed: a "**not useful"** pattern (such as an evident structure visible without any computing or irrelevant to the problem under consideration), a "**useful"** pattern, which provides instant, interesting, and workable information, and a "**to interpret"** pattern**,** which cannot be categorized into the two previous groups and, must be to be studied by experts in the field.

The final goal of the overall methodology is to provide intelligence, or recommendations, that will result in concrete policing strategies and operations. A systematic assessment of results obtained is necessary, leading to possible further refinement of the implemented technology or adaptation due to the evolution of the criminal activity under analysis. In the following, the approach is illustrated through the systematic exploitation of illicit drug seized by the police.

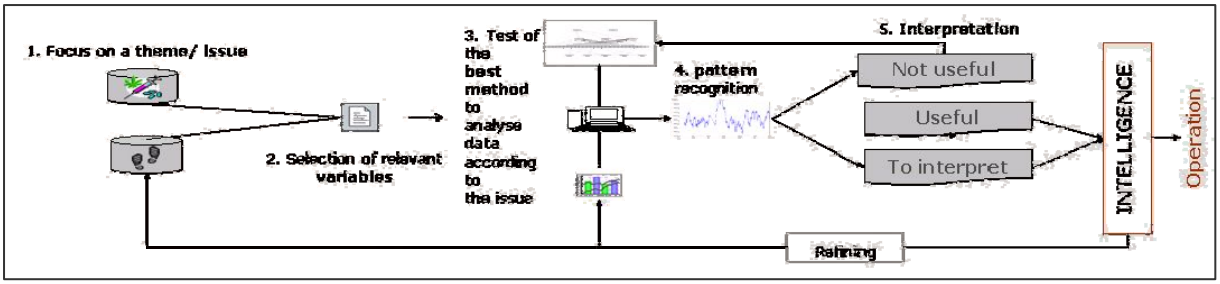

Figure 2 : Forensic sciences application

3. Application to cutting agents found in illicit drug seizures

Drug profiling is an intelligence approach for the systematic use of chemical and physical features extracted through laboratory techniques from illicit drug seized by the police [Guéniat and Esseiva, 2005]. Pattern recognition methods have been systematically tested on set of heroin and cocaine profiles in order to detect possible regularities which may inform on the size and evolution of the traffic [Ratle *et al.* 2006a, Ratle *et al.*, 2006b]. Classical algorithms such as principal component analysis, various clustering and classification algorithms have been successfully applied to heroin datasets.
A new approach based on graph theory and combinatorial analysis has been suggested in order to better exploit cutting agents found in heroin seizures [Terrettaz-Zufferey *et al*. 2006]. The cutting process may occur at different levels of the traffic, but most probably toward the end of the distribution process in order to keep the amount of substance to be conveyed as low as possible. Those agents are thus of particular interest to understand the local traffic network.

The study covers cocaine seizures made in a specific region of Switzerland (the canton of Geneva with a population around 500'000) during one year. The presence or absence of known cutting agents has been systematically detected through chemical analytical laboratory techniques. One single sample may contain simultaneously several different cutting agents. One specific combination of cutting agents may be a marker at different levels of a supply chain. Hence, the dynamic of the occurrence of these combinations can be a good indicator of the evolution of the local market. This process can be naturally modelled through combinatorial analysis and graph theory.

The database created for this study contains the following variables:

- seizure location and time period
- presence/absence of cutting agents
- combinations of cutting agents

The modelling process is explained in figure 3. It focuses on the co-occurrence of cutting agents in the same sample and the persistence of a same combination in time.
.

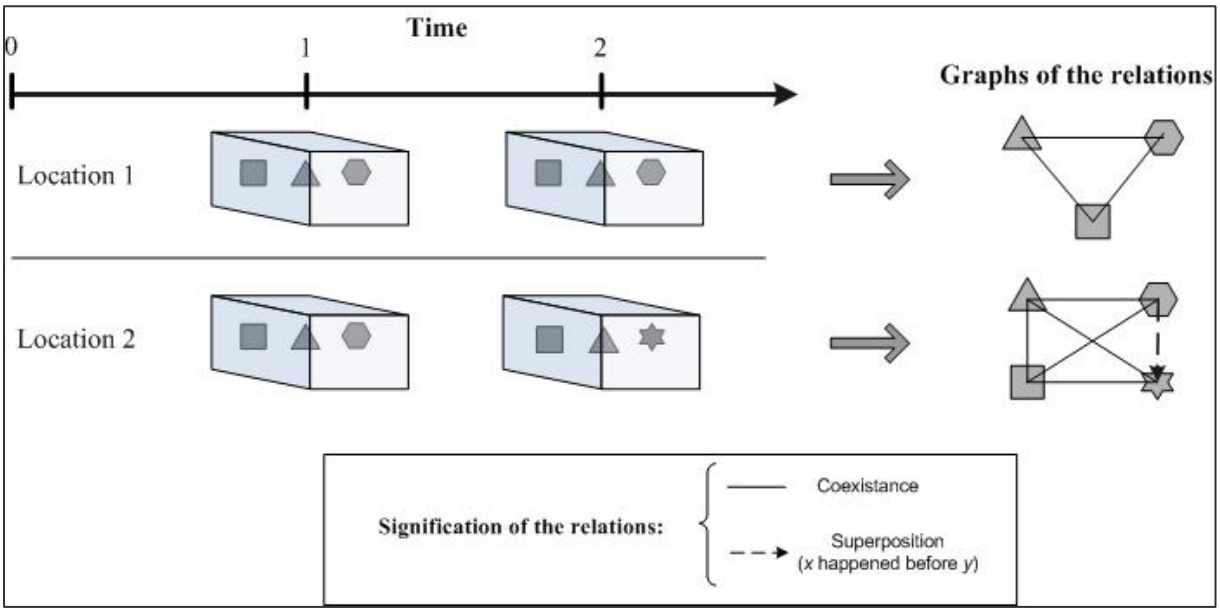

Figure 3 : Construction of the graph. In Location 1, at two different times, the same combination of cutting agents was found. The modelling process leads to the construction of the graph whose nodes represent the cutting agents and edges their co-occurence. In Location 2, the situation is more complex. The two different seizures show the combination of the same two cutting agents, but one differs. The resulting graph can be thought as the superposition of each graph where the same nodes are merged together.

4. Results

A graph of the relationships between cutting agents has been drawn for several localities from border regions to the centre of the canton of Geneva. Geneva is a city that concentrates most of the population of this region (Figure 4). Obviously, graphs resulting from seizures made in the city are more complex than those made at small localities close to the border. It is explained through the more intensive cutting process that prevails inside the country. This reflects the complexity of the local traffic. The seizures made at customs or in regions close to the border give information on the state of cocaine when it is conveyed into Switzerland.

The “life duration” of the combinations of cutting agents is very visible in Figure 5. It is another analytical perspective that shows five groups composed of several combinations. More detailed analyses of the content of the five groups and comparisons with other regions have shown that combinations of groups 1, 2, and 3 occurred only in the canton of Geneva during the same period. Those of groups 4 and 5 were also found in other regions of Switzerland.

Because there is evidence that the structures of the graphs are different from one region to another one, it is possible to consider a combination found in one region during a given time period as a marker for the distribution network. Some kind of subtraction of combinations found at the border (the state in which it is conveyed into Switzerland) may also accentuate these differences.

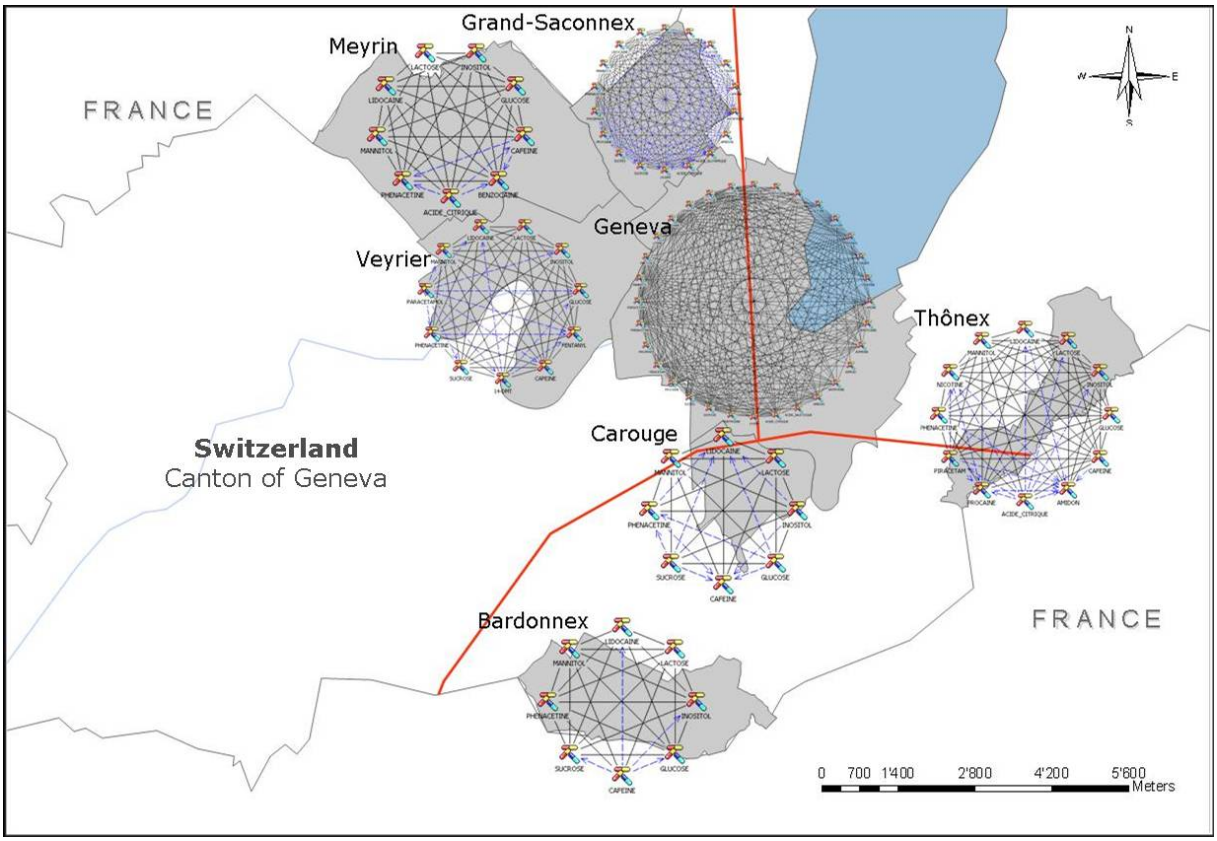

Figure 4: graphs built from cutting agents found in cocaine seized in different areas of the region under consideration

5. Synthesis

Data mining is generally considered as the extraction of useful knowledge from large volume of data. This application is a demonstration that useful patterns can be extract from criminal data after structured preparation of data and by focusing on *a priori* promising and well delineated aspects of a criminal problem where there is evidence that useful patterns can be found: cutting agents are potentially good markers for understanding the local illicit drug market. Accurate delimitation of problem areas where data mining technologies can be concretely implemented is a long term research project.

The interpretation process has established the usefulness of most of the detected patterns, which validates the global approach. The method has still potential for improvement by considering a wider spatial and time spectrum, and building other types of graphs by providing different meaning to the edges (representation of time, conjoint appearance, etc.) or to integrate some knowledge to change their structures (for instance develop specific representations for graphs representing seizures made at the border).

This last point emphasises the role of experts in the field with the overall methodology. Knowledge about the characteristics of forensic case data and the general dynamic of illicit drugs in Switzerland must be integrated in the interpretation of detected patterns.

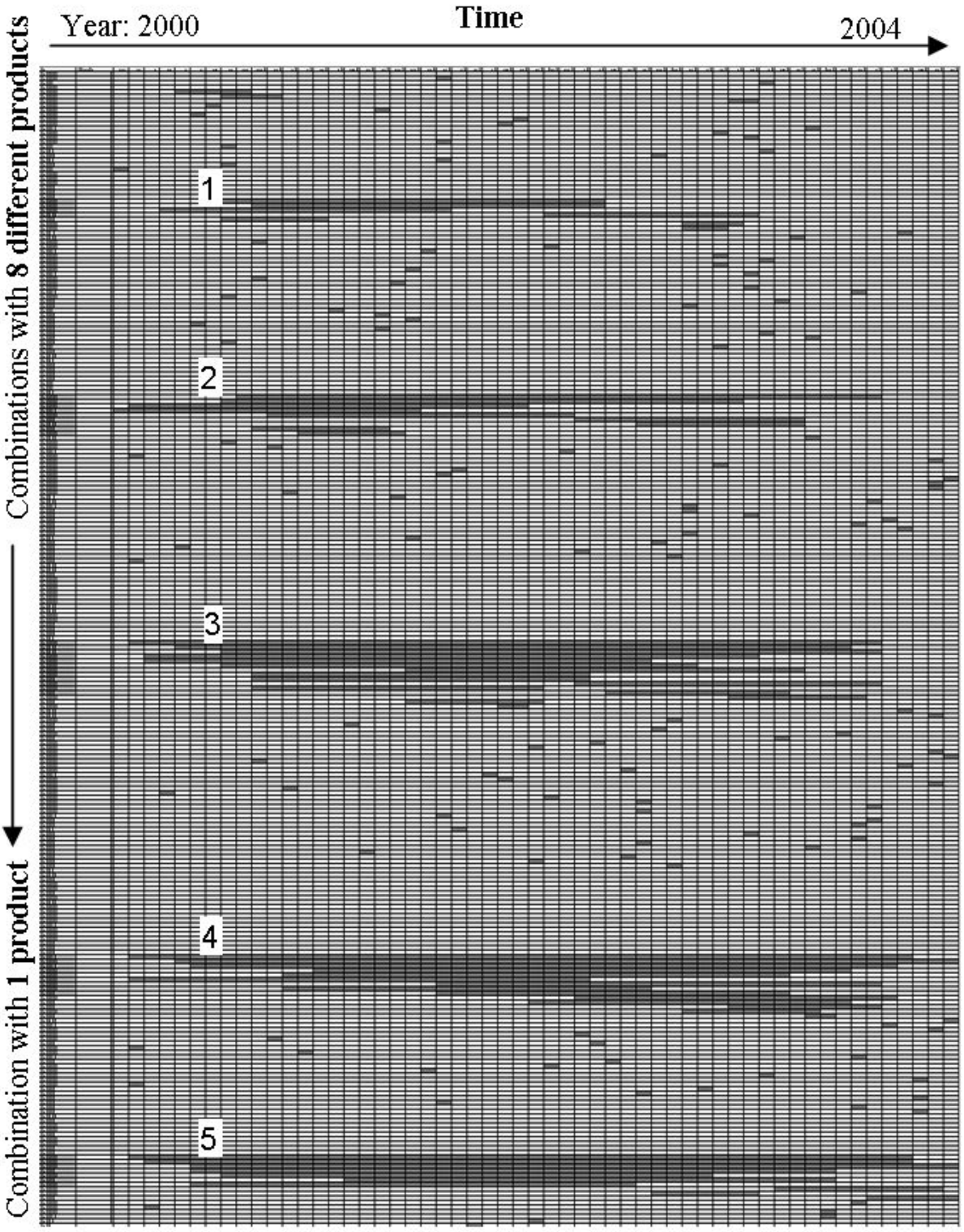

Figure 4 : Persistency in time of combinations in a given place. Each block represents one month on the x-axis and one combination on the y-axis. Combinations are presented ranging from the greatest number of products (8) found in the combination (top) to the least number of products (1, bottom). One can readily appreciate the persistence of a given combination in time.

6. Conclusions

Data mining techniques have a large potential in crime analysis and forensic intelligence. Criminal behaviours can hardly be universally modelled, present a broad diversity, and constantly evolve over time. New modus operandi, drug manufacture processes, or other

novelties may appear as a reaction to policing strategies or social changes such as those caused by the intensive use of new technologies. The development of methodologies including recognition of patterns through data mining technologies could help to adapt to those changes by a detection of new or not yet understood criminal activities. This kind of analysis goes beyond the usual attitude to strictly rely on previous experiences or models developed on limited statistical foundations that constrain the thinking process and prevent in some manner the detection of novelties. However, it is also reasonable to "think small" by limiting at first their applications to realistic dataset that reflect delineated criminal problems. This will help to better understand how to derive useful tools for crime analysis and to distribute efforts between human experts (crime analysts) and the computer.
In this perspective, the exploitation of cutting agents found in cocaine seizures has proved to be a good field of application for combinatorial approach and graph theory. This experience helps imagine the different possibilities of developing principles for crime analysis methodologies that efficiently integrate data mining technologies.

References

GUENIAT, O. AND ESSEIVA, P. Le profilage de l'héroïne et de la cocaïne. Une méthodologie moderne de lutte contre le trafic illicite, *Presses Polytechniques et Universitaires Romandes*, **2005**

RATLE, F., TERRETTAZ-ZUFFEREY, A.-L., KANEVSKI, M., ESSEIVA, P., RIBAUX, O., Pattern analysis in illicit heroin seizures: a novel application of machine learning algorithms, *European Symposium on Artificial Neural Networks*, Bruges, **2006a**

RATLE, F., TERRETTAZ-ZUFFEREY, A.-L., KANEVSKI, M., ESSEIVA, P., RIBAUX, O., Learning Manifolds in Forensic Data, *International Conference on Artificial Neural Networks*, Athens, **2006b**

TERRETTAZ-ZUFFEREY, A.-L. AND RATLE, F. AND RIBAUX, O. AND ESSEIVA, P. AND KANEVSKI, M., attern Detection in Forensic Case Data Using Graph Theory: Application to Heroin Cutting Agents, *Forensic Sci. Int.,* **2006,** In Press