UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

A Knowledge Extraction Framework for Crime Analysis
Unsupervised Methods in Uncertain Environments

PhD Thesis submitted to the Faculty of Economics and Business

Information Management Institute

University of Neuchatel

For the degree of PhD in Computer Science

by

Fabrizio ALBERTETTI

Accepted by the dissertation committee:

Prof. Kilian Stoffel, University of Neuchatel, thesis director

Prof. Olivier Ribaux, University of Lausanne

Prof. Thomas Studer, University of Bern

Dr. Paul Cotofrei, University of Neuchatel

Prof. Catalin Starica, University of Neuchatel

Defended on December 18, 2015

Faculté des Sciences Economiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
www.unine.ch/seco

LETTRES ET SCIENCES HUMAINES
SCIENCES
DROIT
SCIENCES ECONOMIQUES
THEOLOGIE

UNIVERSITÉ DE
NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

A Knowledge Extraction Framework for Crime Analysis :

Unsupervised Methods in Uncertain Environments

Fabrizio ALBERTETTI

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Kilian Stoffel (directeur de thèse, Université de Neuchâtel)
Prof. Catalin Starica (président du jury, Université de Neuchâtel)
Dr. Paul Cotofrei (Université de Neuchâtel)
Prof. Olivier Ribaux (Université de Lausanne)
Prof. Thomas Studer (Université de Berne)

Autorise l'impression de la présente thèse.

Neuchâtel, le 16 février 2016

La doyenne

Carolina Salva

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

A ma mère

Executive Summary

This doctoral thesis investigates the role of knowledge extraction methods in the analysis of crime as an interdisciplinary project, with a focus on unsupervised methods dealing with the uncertain aspects that are intrinsic to the crime environment.

In a context where data generated from criminal activities are increasingly available due to the evolution of technology, the use of automated methods to create value from these data becomes a necessity. These analytic methods require a specific design with regard to the nature of the data they deal with, mostly gathered from crime scenes. Crime analysts desperately need such methods to be better informed and efficient in the perpetual struggle against crime. However, their choices in terms of range and availability are very limited.

A framework delineating and explaining the role of knowledge extraction methods for crime analysis is provided. This framework addresses a particular challenge: developing unsupervised data mining methods dealing with the uncertainty of crime data.

Three approaches are developed to confront this challenge. (1) How to structure and represent crime data to fully exploit the potential of revealing knowledge with further analyses? (2) What is the appropriate method to analyze links between crimes that can deal with both qualitative and quantitative crime data? And (3) what is the appropriate method to help crime analysts to flexibly and understandably detect changes in crime trends?

The significance of this interdisciplinary research can be summarized in two points: it clarifies and delineates the role of data mining in crime analysis, by giving some insights into its applicability in this particular environment; and it makes easier the extraction of knowledge by the use of the proposed domain-driven methods.

Keywords: data mining, domain-driven data mining, knowledge extraction, fuzzy logic, unsupervised methods, soft computing, time series analysis, hybrid methods, crime analysis, crime intelligence, computational forensics, change points detection, trends detection, crime linkage

Résumé

Cette thèse de doctorat investigue le rôle des méthodes d'extraction de connaissances dans l'analyse criminelle en tant que projet interdisciplinaire, avec une orientation sur les méthodes non supervisées traitant les aspects d'incertitude qui sont intrinsèques à l'environnement du crime.

Dans un contexte où les données générées par les activités criminelles sont de plus en plus disponibles grâce à l'évolution des technologies, l'utilisation de méthodes automatisées pour créer de la valeur à partir de ces données devient une nécessité. Ces méthodes d'analyse requièrent une conception spécifique selon la nature des données qu'elles traitent, principalement collectées à partir de scènes de crimes. Les analystes criminels ont désespérément besoin de telles méthodes pour être mieux informés et efficients dans la lutte perpétuelle contre le crime. Cependant, leurs choix en termes d'étendue et de disponibilité sont très limités.

Un framework qui délimite et explique le rôle des méthodes d'extraction de connaissance pour l'analyse criminelle est proposé. Ce framework adresse un défi particulier : développer des méthodes de data mining non supervisées qui permettent de traiter l'incertitude des données criminelles.

Trois approches sont développées pour confronter ce défi. (1) Comment structurer et représenter des données criminelles pour exploiter pleinement leur potentiel à révéler des connaissances par la conduite d'autres analyses ? (2) Quelle est la méthode appropriée d'analyse de liens entre les crimes qui prenne en compte des données à la fois quantitatives et qualitatives? Et (3) quelle est la méthode appropriée pour aider les analystes criminels à détecter des changements dans des tendances criminelles d'une manière flexible et compréhensible ?

L'importance de cette recherche interdisciplinaire peut être résumée en deux points : elle clarifie et délimite le rôle du data mining dans l'analyse criminelle, en fournissant une perspective sur son applicabilité dans cet environnement particulier ; et elle facilite l'extraction de connaissances par l'utilisation des méthodes proposée guidées par le métier.

Mots-clés: exploration de données, exploration de données guidée par le domaine, extraction de connaissances, logique floue, méthodes non supervisées, calcul souple, analyse de séries temporelles, méthodes hybrides, analyse criminelle, intelligence criminelle, computational forensics, détection de points de changement, détection de tendances, analyse de liens entre les crimes

Preface

After completion of studies, the difficult choice of the next career step surely determines a big part of one's life. Deciding to start a thesis after receiving my Master's degree was obviously one of these choices. During these four years of my PhD, many people have asked the same question: *Why starting a PhD thesis?* I have to admit I have not found a "scientific" nor a sensible answer yet, even at the end of this "devoted to science" period. However, the answer I used to give has always been the same: "If you like what you do then you don't need to put this question." Obviously there are many more reasons, but it seems to be the most important one. Upon completion of this essay, I can pretend that many scientists are driven by passion. And pretend further that if you are reading this text you are one of them —excepted for my family and my friends. I hope reading this dissertation will give you as much pleasure as I had producing it.

Both crime analysts and computer scientists will surely find an interest in reading this thesis. The preliminary part gives an overview of both disciplines, which should give the unfamiliar readers to one field or another an insight into the combined problematics. The second part describes particular methods that contributes to the advancement in crime analysis by the completion of specific tasks and in computer science by the investigation of new means to analyze information.

Making this manuscript error-free is left as an exercise for the reader.

Acknowledgements

The first words go to my supervisor, Kilian Stoffel, for the opportunity he gave me and for his unconditional humility. For that I will be eternally grateful.

I would also like to thank Paul Cotofrei and Abdelkader Belkoniene, for valuable advice under any circumstances.

I am also grateful to our partners from the Institut de Police Scientifique (Université de Lausanne), especially Olivier Ribaux and Lionel Grossrieder, but also Damien Dessimoz and Sylvain Ioset for their support.

For the social activities around the thesis, the (so) many sport sessions, and for an authentic friendship, Luciano Lopez.

Last but not least, the Swiss National Science Foundation (SNSF) for having funded this interdisciplinary project (No. 135236 and 156287). The SNSF provides great opportunities to researchers and we will always need their support.

The Contents

This doctoral thesis investigates the challenges of knowledge extraction in crime analysis. It is split into 8 chapters, divided into 2 parts.

PART I - PRELIMINARIES

Chapter 1 – A General Introduction presents an overview of the topic as a whole with the introduction of the proposed framework. It states the motivation and the main challenges of this research, and also briefly presents the three dissertations to address these challenges.

Chapter 2 – The Crime Analyst's Perspective: from Traces to Patterns gives some preliminary concepts in the field of crime analysis. The process starting from the gathering of traces at a crime scene to the discovery of patterns is presented.

Chapter 3 – The Computer Scientist's Perspective: from Data to Knowledge gives a complementary view with some literature in computer science. The main activities involved in the transformation of data into knowledge are tackled.

PART II - THE THREE ESSAYS

Chapter 4 – An Overview of the Three Essays describes the three selected essays and unfolds the search process that lead to these contributions.

Chapter 5 – Structuring Crime Data: from Police Reports to Data Marts details the first essay about crime data structuring. This method lays the ground for fully exploiting the potential of data to reveal knowledge with further analyses.

Chapter 6 – A Method for Crime Linkage details the second essay. A method is proposed to address the specific issue of crime linkage (also referred to as link analysis) with respect to the framework.

Chapter 7 – A Method for Detecting Changes in Crime Trends details the third essay. A method is proposed to address the specific issue of change detection in crime trends (dealt with change points detection) with respect to the framework.

Chapter 8 – Conclusions draws some general conclusions about the framework and suggests some future work.

Neuchatel, December 2015 *Fabrizio Albertetti*

Contents

List of Tables

List of Figures

Part I

PRELIMINARIES

"Give me six hours to chop down a tree and I will spend the first four sharpening the axe."

—Abraham Lincoln (1809 – 1865)

1

A General Introduction

This chapter presents an overview of the topic as a whole and introduces the proposed framework. It states the motivation and the challenges, and also briefly introduces the three selected essays.

1.1 Motivation

Over the last decades, the digital era[1]has been profoundly changing our day-to-day life. The constant development of new computer-related technologies and their increasing adoption rate brought transformations to the social, cultural, and legal aspects of our environment creating a knowledge-based society, where information value is more than substantial. Examples of actors trying to face the evolution race are numerous: governments have been tentatively embracing digital identities; problems related to cyberspace and cybercrime have generated new pieces of legislation; new governance models are being adapted by enterprises to include digital risks and their corresponding insurances; home appliances are being monitored and controlled through smart phones; every single Web query is being stored and analyzed; and digital calls are being used as almost exclusive means of communication (to name only a few specific examples). All of these new digital-related activities lead to a constant increase of data creation, most of the time stored, representing new opportunities as well as new risks for our society.

When considering criminal activities, the number of events being collected is pretty impressive[2]: more than $1,000$ criminal offences against property are registered every day in Switzerland, and, only for the year 2014, a total of $52,338$ were identified as *burglaries* among these. Such figures reveal the number of cases that police units, crime analysts, and legal representatives have to deal with every day. To effectively address the digitization of these cases, only automated methods can support the work of law enforcers. The objectives achieved by these methods can be numerous, e.g., solving more cases, providing a deeper insight into crime phenomena, simplifying operational and administrative tasks, or even to some extent predicting crimes.

In the domain of crime analysis, which is considered by Boba (2009) as the

> "systematic study of crime and disorder problems as well as other police-related issues —including socio-demographic, spatial, and temporal factors— to assist the police in criminal apprehension, crime and disorder reduction, crime prevention, and evaluation",

the use of automated methods is particularly noteworthy with the emergence

[1] "The Digital Era can be seen as the development of an evolutionary system in which knowledge turnover is not only very high, but also increasingly out of the control of humans, making it a time in which our lives become more difficult to manage." (Shepherd 2004)

[2] Source: Statistique policière de la criminalité (SPC), OFS. Please note that a single criminal event can generate several offences codes.

of so-called *predictive policing* models. In this paradigm, a better assistance to the police and a deeper understanding of crimes is advocated. This proactive policing model is targeted by the creation and the use of timely intelligence with the objective of reducing crime. New and disruptive analysis methods devised by both crime analysts and scientists are expected to fulfill this vision, contributing to the advancement of the discipline. For a successful approach, it becomes critical to support and to make use of the products of the digital era. Indeed, the ever-increasing development of technologies has been more and more supporting human activities, and as a corollary also increasing everyone's traceability. An enormous effort must be made to include the product of these new traces within the automated reasoning mechanisms to facilitate the work of the crime analysts. However, it implies that more complex cases have to be handled by police agencies, not to mention the need to face privacy and legal constraints, or the new skills that crime analysts need to acquire to master these new tools.

Much effort has been made in the literature to develop data mining/knowledge extraction[3] methods for the analysis of crimes, but very few research studies provide a unifying framework describing how these methods should be used and how they are related. Ribaux et al. (Ribaux and Margot 1999; Ribaux, Girod, et al. 2003) placed an emphasis on the lack of integrated approaches concerning crime analysis solutions.

Nonetheless, several attempts to adopt a more complete view have been made. As a notable example, an emerging interdisciplinary domain, namely Computational Forensics (CF) (see Franke and S. Srihari 2008), provides a partial answer by deriving mathematical, statistical and computational science methods to investigate forensic issues, with the main goal of discovery and advancement of forensic knowledge. However, this focus is too narrow to fully address crime analysis issues, whereby only forensics-related problems leading to forensic knowledge are targeted. Furthermore, computational methods must be framed by integrating theories derived from forensic science, criminology and crime analysis. Other studies have also tried to provide a holistic answer to this general issue in crime analysis. For example, the well-known studies such as the RECAP project (Brown 1998), the Regional Crime Analysis Program providing both data fusion and data analysis for crime analysts; the COPLINK project[4]defining an extensive analytical environment (H. Chen, Schroeder, et al. 2003; H. Chen, W. Chung, et al. 2004); or a doctoral thesis (Adderley 2007) aiming at defining the role of data mining

[3] In the literature, *knowledge extraction* and *data mining* are often used to denote the same concept. In this document, both words will be used interchangeably.

methods in operational policing.

1.2 Contribution

In this dissertation, a knowledge extraction framework for crime analysis dealing with uncertain environments is proposed. The goal of this framework is the overall advancement of the crime analysis discipline through the discovery of crime-related knowledge with the means of computational methods, within an interdisciplinary and proactive approach.

Stemming from a collaboration between crime analysis experts and computer scientists, several disciplines are put together to provide an integrated approach. Computational science, forensic science, crime analysis, and criminology are combined to deal with crime-related issues. Put differently, *this research seeks to deliver a framework for developing relevant analytical methods with the purpose of extracting knowledge from crime data, systematically motivated and validated with crime theories.* This latter point is essential to provide both valid and sound methods, in contrast to isolated solutions with weak criminological groundings.

This framework does not seek to be exhaustive as it would be in a typical review assessing the systematic applicability of *every* mining method to crime analysis. It rather intends to investigate the appropriate context for developing knowledge extraction methods in the specific field of crime analysis. In practical terms, this is achieved by providing methods to crime analysts. These methods aim to achieve specific tasks with intelligent computational-related tools, and ultimately to extract new knowledge (Fig. 1.1).

The need for such a framework can be put forward by the following arguments: (a) there is an increasing awareness that a truly interdisciplinary and proactive approach integrating both computational and crime analysis considerations is needed; (b) crime analysis issues are practical problems grounded in reality of experts' domain and need relevant solutions to address them in order to produce useful and timely intelligence; (c) only sophisticated computational

[4] A research project integrating several databases for a law enforcement Intranet. Carried out in a first time Schroeder and Dept 2001 by the Tucson Police Department and the National Institute of Justice (NIJ) in 1997, the COPLINK project was taken over by IBM and adopted by many police units in USA. Well documented and with many case studies Hauck and H. Chen 1999; H. Chen, Schroeder, et al. 2003; H. Chen, W. Chung, et al. 2004; Atabakhsh et al. 2001, the scientific part is still provided by the National Criminal Justice Reference Service (NCJRS).

methods integrating crime-related knowledge are able to produce interesting outcomes with a high degree of acceptance by crime analysis experts; and (d) many existing solutions do not investigate automated methods with a holistic approach: almost no guidance is given about how data should be structured prior its analysis and explanations are simply lacking on how results should be integrated and interpreted.

To develop this framework, a partnership has been created between the School of Criminal Sciences[5] and the Information Management Institute[6]. The former provides the crime analysis expertise and the latter the computational expertise, working as an interdisciplinary team.[7]

1.3 Challenges and Research Tracks

The first challenge is to deal with crime data. Concerning any aspect about crime that can be collected starting from the crime scene, crime data can serve many purposes, e.g., crime investigation, crime detection, or court evidence. Hidden in a mostly chaotic environment, pieces of evidence are "mainly fragmentary, latent, and inevitably distorted" (Franke and S. Srihari 2008). As a consequence, the data gathered are of poor quality: reasoning is performed on the basis of uncertainties, partial knowledge and conjectures. Furthermore, many analysis tasks have unclear objectives, meaning that unlabeled data is presented to the learning process. Therefore, only specific methods, that is, unsupervised and handling uncertain environments, can make up the proposed artifact.

Another main challenge is to find appropriate means to integrate and to represent domain knowledge for reasoning (forensic knowledge, legal knowledge, business processes, or any other piece of knowledge assuming some more intelligent/oriented results). Mostly qualitative, this kind of knowledge requires flexible extraction methods, where most mining methods deal with quantitative data only. The integration of intelligence should enhance both inference capabilities and human comprehensibility, achieved either by adjusting existing models to the experts' rationale, or by adapting the mechanisms

[5]Institut de Police Scientifique, Ecole des Sciences Criminelles, Université de Lausanne

[6]Institut du Management de l'Information, Faculté des Sciences Economiques, Université de Neuchâtel

[7]One of the project objective is the production of two complementary PhD theses: one focusing on crime analysis related issues; and the other, the present document, focusing on computer science related issues.

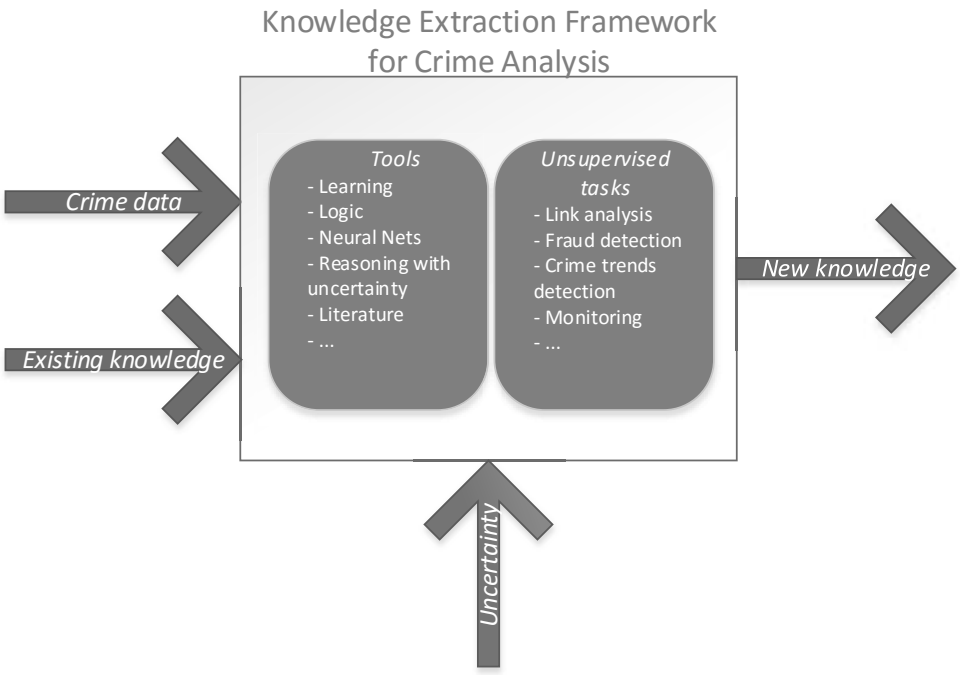


Figure 1.1: The proposed knowledge extraction framework for crime analysis. Starting from *crime data* and *existing knowledge*, intelligent *tools* are used to achieve *unsupervised tasks*, dealing with *uncertainty* stemming from the environment.

with a formalism capturing the impreciseness and vagueness of crime data.

The development of the proposed framework with the aforementioned challenges raises several questions, both practical and theoretical. To put the matter in a nutshell, here are some of these:

- Can knowledge extraction methods be put within the collection of existing crime analysis approaches?
- What is the appropriate theory on which reasoning can be performed with uncertainty and vagueness?
- What is the appropriate formalism to represent intelligence?
- How can intelligence be included within the global process of crime analysis?
- What are the knowledge extraction methods that can deal with *qualita-*

tive data in crime analysis?

- What are the *unsupervised* knowledge extraction methods that can deal with crime data?
- What are the mechanisms allowing the execution of a crime analysis process to influence/fine-tune data mining algorithms?
- How to automatically assess the correctness and the quality of the extracted knowledge, regarding domain knowledge?

As an attempt to address some of these questions, the three following approaches are considered in this dissertation, each with a dedicated chapter:

1. Find an appropriate means to structure crime data in order to fully exploit the potential of knowledge extraction methods.
2. Devise a crime linkage (also known as link analysis) method based on both quantitative and qualitative features.
3. Devise a flexible method to help crime analysts detecting changes in crime trends.

The first contribution describes a crime data structuring method based on data warehousing concepts. Uncertainty of crime data is partly mitigated. For example, instead of mining a poor set of events with various features such as a police log, the same information can be represented by using some more appropriate structures/dimensions (e.g., aggregating events into time series or histograms). Furthermore, the information value can be increased by adding some information (e.g., some meta information related to the objective). It lays the ground to conducting analyses more efficiently.

The second contribution provides a crime linkage method that can consider both qualitative and quantitative data. It is based on a multi-criteria decision making to express the preferences of the analyst in regard to the importance of the features.

The third contribution delivers a change points detection method for crime-related time series, dealing with uncertainty. Fuzzy logic enables the development of flexible approaches and gives the ability to crime analysts to express their inputs with linguistic variables.

Besides, these three approaches are implemented and validated in accordance with the proposed framework, always with the objective of extracting knowledge with the means of data mining techniques. Last but not least, one should keep in mind that the primary focus of this dissertation is to contribute to the

computational part, although an interdisciplinary approach is systematically pushed forward.

More broadly, this framework clarifies the role of data mining in crime analysis and can be used to develop new methods more efficiently. Indeed, a very limited number of studies address this matter with a comprehensive approach.

The remainder of this document is structured as follows: Chap. 2 describes the perspective of the crime analyst in general. Similarly, the perspective of a computer scientist is given in Chap. 3. Chapter 4 introduces the details of the three contributions and explains the search process that lead to these outcomes. Chapter 5 entails the first contribution with a method for structuring crime data. Chapter 6 describes the second contribution with a method for crime linkage. Chapter 7 describes the third contribution with a method for change points detection. Finally, some general conclusions and future work are presented in Chap. 8.

"Give a man a fish and you feed him for a day. Teach a man to fish and you feed him for a lifetime."

—Chinese proverb

2

The Crime Analyst's Perspective: from Traces to Patterns

In this chapter, we attempt to understand the path starting from the reading of a set of crime cases leading to the discovery of new patterns or knowledge. First, the concept of crime analysis is introduced. Then, we present how criminal approaches can explain a certain amount of rationality in crimes by comparing the opportunities and risks that criminals may face. Understanding that the occurrence of a crime is not completely random, but based on the environment and opportunities, helps us to leverage heuristics,leading to a more intelligent use of data mining techniques. Finally, the process in which the production of intelligence takes place is presented.

2.1 What is Crime Analysis?

As already mentioned, crime analysis can be described as the "systematic study of crime and disorder problems [...]" (Boba 2009). Alternatively, Gottlieb, Sheldon, and Raj (1994) described crime analysis as

> "a set of systematic analytical processes directed at providing timely and pertinent information relative to crime patterns and trend correlations to assist the operational and administrative personnel in planning the deployment of resources for the prevention and suppression of criminal activities, aiding the investigative process, and increasing apprehensions and the clearance of cases."

In an attempt to clarify what crime analysis is and to provide a common definition for practitioners, the International Association of Crime Analysts (IACA) has recently proposed the following formulation (International Association of Crime Analysts (IACA) 2014):

> "Crime analysis is a profession and process in which a set of quantitative and qualitative techniques are used to analyze data valuable to police agencies and their communities. It includes the analysis of crime and criminals, crime victims, disorder, quality of life issues, traffic issues, and internal police operations, and its results support criminal investigation and prosecution, patrol activities, crime prevention and reduction strategies, problem solving, and the evaluation of police efforts."

I must emphasize that the term "crime" in *crime analysis* is too often — wrongly— interpreted as a restricted term referring to crimes only; it actually encompasses much more:

- the crimes themselves
- disorders
- incidents
- quality-of-life issues
- traffic collision events
- and any other piece of information relating to police agencies

However, the analysis of *evidence* (including DNA) is excluded from their definition.

Four major categories of crime analysis are recognized by the IACA (Fig. 2.1), hereafter ordered from specific to general:

a) crime intelligence analysis;

b) tactical crime analysis;

c) strategic crime analysis; and

d) administrative crime analysis.

The first category, crime intelligence analysis, focuses on data about people involved in crimes. We need to emphasize here that the term intelligence refers to information about authors, but not to high-level information about crime in general (in the sense of knowledge). The objective of this category is therefore to deter criminal activities by understanding the context in which they lie. The main criminal activities involved are repeat offenders, repeat victims, and serial crimes. These activities may use police logs as a starting point and are subject to be conducted on a daily basis.

The second, tactical crime analysis, investigates space, time, offenders, victims, and modus operandis stemming from police databases to develop short-term recommendations. Crime patterns and crime linkage are typical examples of analyses conducted in this category to define investigation priorities and resource allocation.

Strategic crime analysis uses numerous information sources in the development of policies and prevention techniques in a long-term approach. Trend analysis and hot spot analysis are part of this category.

The last category, administrative crime analysis, concerns the general organization of police agencies in terms of communication, budget, statistics, employment, etc.

Considering our general objective of knowledge extraction, strategic crime analysis would be the most likely category to contributing to the advancement of the discipline, due to its abstraction level. Indeed, crime intelligence and operational crime analysis are less subject to the advancement but more focused on solving specific crimes and identifying authors with limited use of "intelligence" (intelligence here refers to interpreted data from its data mining definition).

However, in practice, many resources are devoted to the first and second category. Trend analysis, crime pattern analysis, or linking known offenders to past crimes are all examples of common tasks done within crime analysis agencies. It is particularly true regarding our partners, that is, the crime

Source: International Association of Crime Analysts (IACA) (2014)

Figure 2.1: The types of crime analysis in four dimensions

analysis department from the Police du Canton de Vaud and the School of Criminal Sciences from the University of Lausanne.

2.1.1 Models of Crime Analysis

Many models have attempted to conceptualize crime analysis. As instance, let us start with the conceptual *3i* model (*interpret*, *influence*, and *impact*), devised by Ratcliffe (Figure 2.2). In his revised version (Ratcliffe 2008), he describes the role of the prevention within the crime analysis process, by emphasizing on the importance of crime analysts in law enforcement agencies. It puts them as active protagonists instead of passive or reactive actors. Interactions are described as such: *intelligence analysis* helps to *interpret* the *criminal environment*; *intelligence analysis* has an *influence* on *decision-makers* because they produce and understand forensic intelligence; and *decision-makers* on the other side *impact* the *criminal environment*.

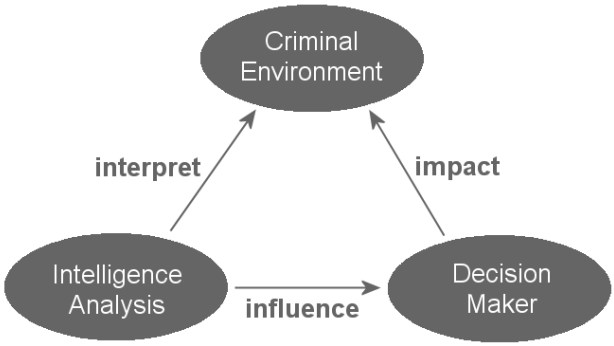


Figure 2.2: The 3i model of crime analysis. The three associations, indicated with arrows, represent the three "i" of the model.

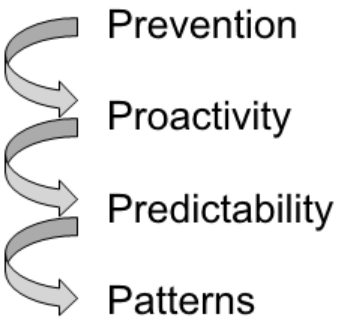


Figure 2.3: The 4P model of crime analysis, a foundation of the intelligence-led policing model

Another model based on intelligence-led policing, the 4P model (Ratcliffe 2009), describes crime analysis with a strong emphasis on prevention (Fig. 2.3). To ensure a preventive approach instead of a reactive approach, the chain reaction is defined as such: the police has to deal with *prevention*. This *prevention* requires *proactivity*, which requires *predictability*, which can be expressed by *patterns*.

This chain advocating crime prevention is at the heart of the *intelligence-led* policing concept. As coined by Ratcliffe (2008), intelligence-led policing is

> "a business model and managerial philosophy where data analysis and crime intelligence are pivotal to an objective, decision-making framework that facilitates crime and problem reduction, disruption and prevention through both strategic management and effective enforcement strategies that target prolific and serious offenders."

2.1.2 Crime Analysis Supporting Crime Investigation

To better understand the role of crime analysis in solving crimes, let us understand the wider process in which crime analysis takes place, that is crime investigation. Crime investigation (also referred to as criminal investigation) is considered to be an applied discipline involving the collection and study of information related to a crime with the objective to identify and charge its author. According to Kind (1994), the crime investigation process —starting from the investigation of a crime and ending up with its trial— can be described in three chapters: (a) the problem to find; (b) the decision to charge; and (c) the problem to prove. The first step consists in the analysis of the criminal event in order to understand what happened and who did it. Then, the presumed offender will be arrested and a case with the related clues is prepared and transmitted to the court. Finally, the investigation can lead to a charge and a trial takes place to assess the guilt of the author. Crime analysis can support all of these steps, mostly by providing useful information helping in making decisions.

2.2 Crime Prevention Approaches

In the late eighties, a paradigm shift from traditional to environmental approaches changed the way to perceive the rationale of crimes. Instead of being criminal-centric, that is considering crime as a result of particular socio-psychological traits and the behavior the offender reveals, this new approach consists in a collection of theories moving the focus onto the *environment* in which a crime occurs. By modifying some aspects of the environment, the opportunities to commit a crime can be reduced.

Most current environmental theories can be illustrated with the idea of *sensitive dependency to initial conditions* (aka the chaos theory), as illustrated by Kuhn (2000). The sensitive dependency to initial conditions is a determinism-based theory suggesting a certain non-linearity and unpredictability of physical events. In this sense, these events are *deterministic in a short term approach*; *but unpredictable in a long term approach.* It excludes the fact that every single aspect has been already defined since the creation of the universe (which would be an absolute deterministic approach); but also excludes the fact that every possibility is given at any time (free will is only partial, being a consequence or a subset of the previous acts, which are themselves a consequence of the previous acts, and so on).

The well-known example of the leaking faucet illustrates the complexity of this (un)predictability: Is it possible to completely predict how the drops will fall, knowing every single element of the system (e.g., the diameter of the faucet, the water throughput, the quantity of sediment within the pipe, the hardness of the water)? As one can imagine, this system does not seem linear, and by modifying —even very slightly— any of these parameters, the space of the new possible states becomes incredibly huge, as if it were the result of an infinitely creative process (or even a random process!) This is what is meant by *sensitive dependency to initial conditions*.

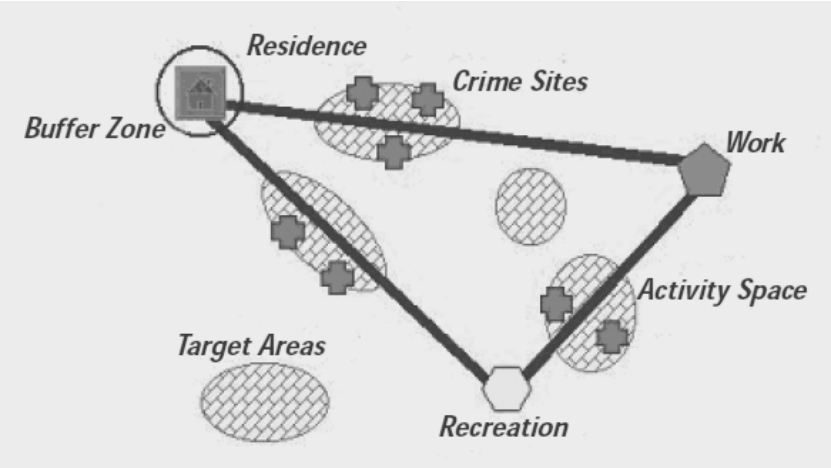
In crime analysis, this can be translated into the proverb "opportunities make the thief", where the change of a tiny element can make a crime to occur, or, similarly, prevent a crime to occur. Many theories have been developed following this trend, such as the routine activity theory (Cohen and Felson 1979), the rational choice perspective (Cornish and R. V. G. Clarke 1975; Cornish and R. V. G. Clarke 1996), or the crime pattern theory (P. L. Brantingham and P. J. Brantingham 1990; P. Brantingham and P. Brantingham 2008). These theories play an important role in policing models where the prevention is central (such as the intelligence-led policing model).

The routine activity theory expresses the necessary conditions for a crime to occur, namely a physical encounter between a motivated author, a suitable target, and the absence or the weakness of a guardian (which can be either human or non-human). By *motivated author*, it is assumed that almost any human can be a motivated author, considering human greed or selfishness as sufficient justifications. The *environment* dimension of a crime is considered as a series of factors that can be acted upon to mitigate crime opportunities.

In the rational choice perspective, the assumption is different and is borrowed from other disciplines. A cost-benefit rationale drives the potential criminal to adapt to the situation with the objective to maximize its utility. For example, if the risk is perceived as too high relating to the outcome, the criminal action will be avoided.

Finally, in the crime pattern theory, explanations are given about how motivated authors meet suitable targets in the spatial dimension. In the offender's activity space, *target areas* denote this idea. These are areas that are close enough to familiar routes but not on the route itself (such as on a park nearby the daily route from home to work). They imply little effort and are therefore favorable situations to commit a crime. In contrast, *buffer areas*, such as the workplace, are avoided (Fig. 2.4).

The main idea of most environmental/situational approaches can be repre-



Source: Rossmo (2000)

Figure 2.4: Crime pattern theory, represented by the offender's activity space (residence, work, and recreation). The offender does not usually commit crimes within the buffer zone of these spaces, but can benefit from the vicinity of target areas (e.g., parking lot) as an opportunity to commit a crime when the three conditions of the crime triangle are met.

sented by the crime triangle analysis (Fig. 2.5). The crime triangle states that by acting on one of the three controllers, a crime can be prevented (the *handler*, the *manager*, and the *victim* are the *controllers* that can minimize the possibility of a crime to occur).

The statistical distribution of crimes also plays a leading part in these approaches, wherein the 80/20 Pareto distribution seems to be also visible (Boba 2009), although measured in more extreme magnitudes. Indeed, on the criminal side, only a minority of offenders are known to commit most of the crimes (Heaton 2000; Wolfgang, Figlio, and Sellin 1987). On the victim side, this statement is also true (Weisel 2005): only a few victims tend to be systematically victimized. And finally, very few areas generate a significant part of the crimes (Sherman, Gartin, and Buerger 1989). To summarize, the salient point of these approaches for our concern is that some crimes do actually follow patterns and can therefore be ultimately predicted.

Source: R. Clarke and Eck (2005)

Figure 2.5: The crime triangle explaining situational approaches with a conjunction of the handler, the manager, and the guardian.

2.3 On the Way to Intelligence

In this section we investigate more closely the product of crime analysis, that is intelligence. We start with the concept of dedicated monitoring processes, then we review the role of crime analysis tasks that are subject to produce intelligence, and finally we attempt to define the term intelligence.

2.3.1 The Desperate Analysts's Need for Automated Methods

Over the last decades, forensic sciences and crime analysis have been adopting the digital era; crime case data have been digitized and their analysis methods automatized as well. Many new methods and tools benefiting from this change have been devised to support analysts in their daily tasks. In forensic science, the analysis of DNA, shoe marks, fingerprints, etc. henceforth benefit from a relatively high level of automation, helping human minds in overcoming their limitations. Previously, for instance, comparing minutiae stemming from a trace of a fingerprint found at crime scene, to, an entire source base containing thousands of entries, was individually performed. Nowadays, mainly thanks to the AFIS technology (Wilson and Woodard 1987) and its evolutions (Maltoni

et al. 2009), a crime analyst unit can rely on an algorithm to almost instantly select a subset of records from a database containing 400 million records. Only then, a single analyst might manually compare the pre-selected subset.

Despite the constant solicited requests for the development of new methods (Giannelli 1998) resulting in a plethora of choices (C. Li 2009), DNA analysis is the only one to be considered as *scientifically rigorous*[1]. Many other studies also comment the quality of available forensic methods in other respects. As instance, Kohn, Eloff, and Olivier (2006), Ribaux, Girod, et al. (2003), H. Chen, W. Chung, et al. (2004), Ribaux, Baylon, Lock, et al. (2010), Ribaux, Baylon, Roux, et al. (2010), and Ribaux, Walsh, and Margot (2006) criticize a partial exhaustiveness of these methods, emphasized by the lack of unifying frameworks. To overcome these limitations, the studies conclude by advocating an interdisciplinary approach by design. Their main argument is the need to gain some insight into multi-domain issues inherent to the real-world. This statement is also valid for crime analysis, where automated methods are more than necessary to support analysts in their daily tasks.

A difficulty in having proven methods can reflect a lack of maturity in the research domain. It is also relevant to notice the discrepancy in the terms currently used to denote computational approaches in forensics: *forensic statistics*, *forensic information technology*, *forensic intelligence*, or *computational forensics* (Franke and S. N. Srihari 2007). These definitions share a similar spectrum, but there is no clear delineated boundary defining how they overlap or not.

To clearly make a distinction between the science itself and its computational approaches, let us see some definitions. *Forensic science*, or *forensics*, is commonly defined as "the methodological application of a broad spectrum of scientific disciplines to answer questions significant to the legal system" (Steen and Blom 2007), or also more generally as "the study of trace evidence, focusing on legal, investigation, intelligence, or prevention related issues". As a new field of forensic science, *computational forensics* (CF) has been recently introduced. CF is defined as the *application of computational methods* aiming

[1] After two years of deliberation summarized by the publication of the *NAS report* (National Research Council 2009), the National Academy's science forensic committee ended up with a troubling picture: "A large part of current forensics practice is skill and art rather than science, and the influences present in a typical law-enforcement setting are not conducive to doing the best science. Also, many of the methods have never been scientifically validated. Among all the classical forensics methods, the committee concluded, only DNA analysis has been shown to be scientifically rigorous. But committee members identified several areas where the greater use of computers and automation could eliminate bias and minimize error."

at solving forensic issues and producing forensic intelligence. This latter is considered as an emerging discipline with quantitative approaches using statistical, mathematical and computational techniques. In this context, it is said that forensic science is *assisted* by algorithms and software, from several areas in computer science. More specifically, CF supports forensic experts in three ways (Franke and S. Srihari 2008):

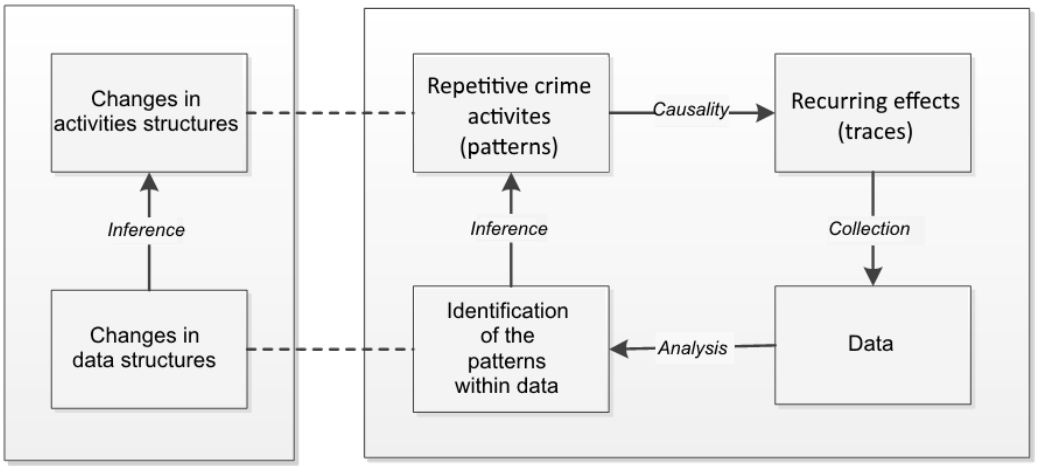
- it provides tools for the human examiner to better analyze evidence by extending human cognitive abilities;
- it supports the forensic discipline by enabling the analysis of large volumes of data which are beyond human cognitive abilities;
- and it can ultimately be used to represent human expert knowledge and to perform automated reasoning and recognition.

2.3.2 Intelligence in Crime Analysis

The refined product of the crime analysis process is known as *intelligence* [2], being the accurate, timely and useful product of logically processing case data (see Ribaux, Walsh, and Margot (2006) and Ribaux, Girod, et al. (2003) for more details). We cannot ignore the link between this definition and the main objective of data mining algorithms, that is knowledge extraction. In the case of crime analysis, intelligence has a limited number of forms it can take: trends, relations, and crime patterns are the three main forms of intelligence.

In Grossrieder, Albertetti, Stoffel, and Ribaux (2013b), we formulate the necessary hypotheses to detect structural changes from criminal activities, which can be interpreted as a prerequisite to find intelligence from a crime data set (Fig. 2.6). Basically, on one hand, crime patterns, trends, or relations can be identified by computational methods only if these structures are reflected within the data; which is the analytical condition. On the other hand, criminal activities produce traces, and these traces need to be reflected within the data; which is the forensic condition.

[2] In crime analysis, "intelligence" is one of the most confounding terms (International Association of Crime Analysts (IACA) 2014). It is sometimes used to refer to a source, sometimes to a product of analysis. The IACA apply it mostly to people (or offenders) rather than to general crime information. Furthermore, when considered as a product, a distinction is made between knowledge and intelligence. *Intelligence* is a refined product or a subset of *knowledge*, which has the specificity of being actionable (i.e., which can be turned into an action). This latter distinction is not directly drawn in the computer science literature, but *knowledge* itself is generally considered as a non-trivial, actionable, and new product of data mining. In this dissertation, *intelligence* will be used to denote



Source: Translated from Grossrieder, Albertetti, Stoffel, and Ribaux (2013b)

Figure 2.6: Hypotheses on how structural changes are detected from criminal activities.

More broadly, intelligence is both an input and an output of the crime analysis process (or any monitoring process). As stated earlier, crime analysis is the systematic study of crime and disorder problems to assist the police in criminal understanding. Therefore, by augmenting data mining techniques with crime intelligence, reasoning steps are imbued with meaning, context, and links to similar previous experiences. This integration helps to produce a more intelligent, oriented, and learning-based output. Such synergies can only be reached through an interdisciplinary approach.

2.3.3 Monitoring Processes

According to its strict definition, the crime investigation process, aided with crime analysis, generates and uses information. In its extended form, the crime investigation process can lead to the production of intelligence. For this purpose, some dedicated approaches have been proposed, known as *monitoring processes*.

Monitoring processes are an answer to the ever-increasing amount of crime data in analysis for gathering and capitalizing intelligence. They contribute to a deeper understanding of criminal acts and of criminals, and fall within

information/knowledge in its data mining definition —that is, as interpreted data from the KDD process—, and considered as a product of analysis (unless otherwise stated).

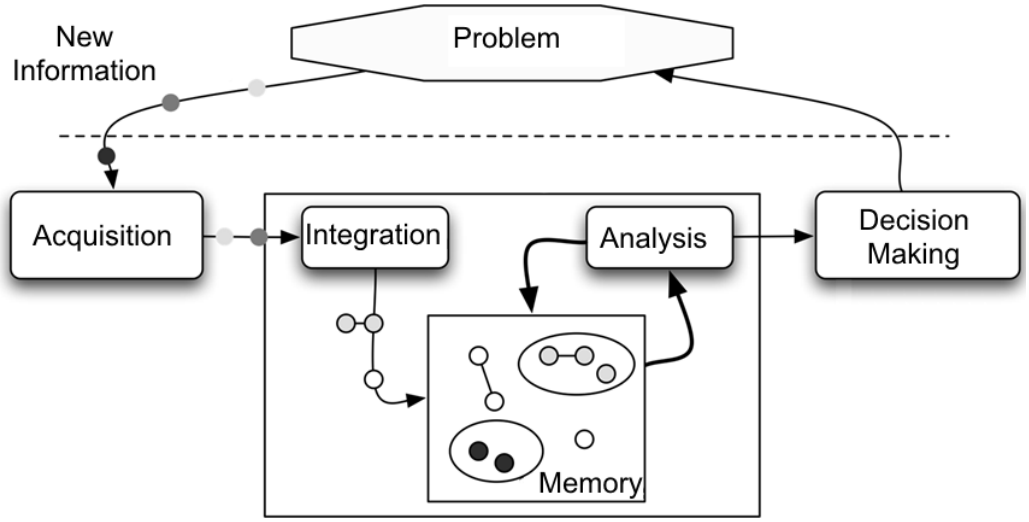


Figure 2.7: The operational monitoring process

Source: Translated from Ribaux, Genessay, and Margot (2011)

the scope of intelligence-led policing models. An example of these monitoring processes is the *operational monitoring process* (Ribaux, Genessay, and Margot 2011; Morelato et al. 2014). This process considers a *systematic* analysis of criminal activities with the objective of discovering and detecting serial crime —which can lead to the discovery of patterns— as a continuous process (Fig. 2.7). The product of this process can be used as an operational resource (serving police agencies) or a strategic resource. In the former case, it benefits to a *specific* investigation in the limited range of operational measures. In the latter case, the product helps in the *overall* understanding of crimes and criminals.

2.3.4 Crime Data Mining Techniques

As mentioned above, intelligence can be extracted with the help of data mining techniques. These techniques can use various sources of information to extract intelligence: traces/forensic-related data (e.g., shoe marks, drug profile, DNA); situational data (e.g., location, date); behavioral data (e.g., modus operandi); and relational data (links between cases) from a set of criminal cases.

A framework proposed by H. Chen, W. Chung, et al. (2004) identifies the relationships between crime types and crime data mining techniques used in

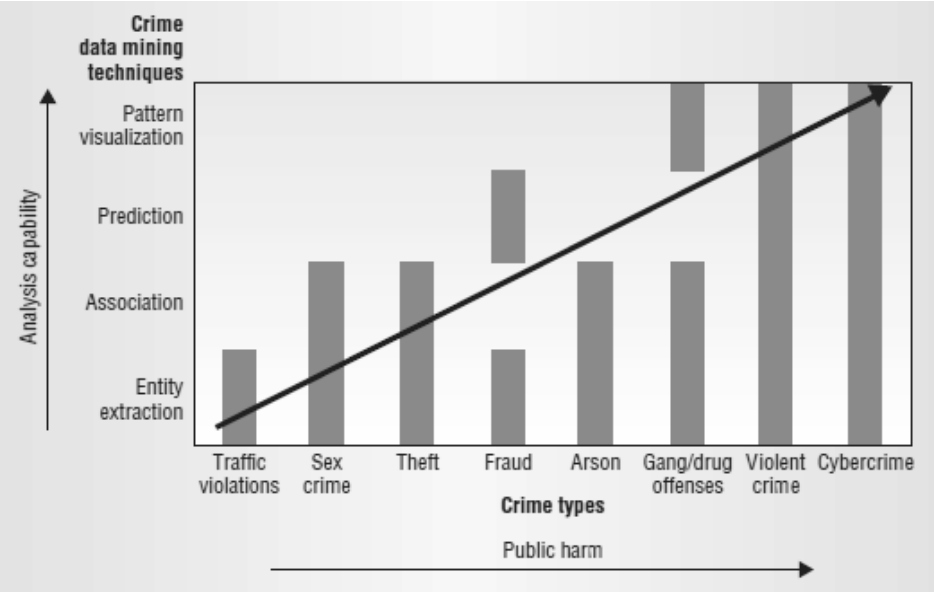


Figure 2.8: H. Chen, W. Chung, et al. (2004) crime data mining framework

Source: H. Chen, W. Chung, et al. (2004)

crime analysis (Fig. 2.8).

The four data mining techniques supporting investigators are entity extraction, association, prediction, and pattern visualization, ordered by their analytical capabilities. These tasks are mapped to various crime types with respect to the public harm they represent. For example, neural networks can be used as prediction for frauds related to credit cards, and clustering can be used for detecting hot spots generated by gang offenses.

Of course, this framework is relatively limited and is pretty rough on its categorization of data mining techniques, but it has the advantage to give an overview of which crime data mining technique is suitable for which crime type.

"An investment in knowledge pays the best interest."

—Benjamin Franklin (1706 – 1790)

3

The Computer Scientist's Perspective: from Data to Knowledge

In this chapter, an overview of some means to prepare and analyze crime data is proposed. The main objectives are (a) to understand that data stemming from traces that are then gathered in police reports should not be analyzed and that a more appropriate structure is required prior conducting analyses. And (b) to get a good grasp on the appropriate methods to transform and analyze these data, that is, what are the steps to transform *data* into *knowledge*.

3.1 Storage Architectures

3.1.1 Transactional/Operational Architecture

A transactional/operational architecture, dealing with day-to-day transactions, uses data stored in a database supported by a database management system (DBMS). Optimized for operational purposes, this kind of environment is called on-line transaction processing. Not really designed for long term analyses nor strategic analysis purposes, data stored within these systems are supposedly structured and normalized to avoid redundancy and to preserve consistency (e.g., in third normal form or Boyce-Codd normal form). Conceptual entity-relationship (E-R) or relational models (contrasted with multidimensional models) are used to design these systems.

3.1.2 Data Warehousing Architecture

Whereas operational systems support business processes, data warehouse (DW) systems support analytical decision processes. A widely accepted definition of a data warehouse is "a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making" (Inmon 2002). We slightly contrast this statement by claiming that a DW is not tightly coupled to decision making and is not subject-oriented; these features are transferred to another level, namely the data marts. Therefore, the data warehouse is not directly serving or supporting a decision support system, but the data mart is. Actually, the general purpose of such data architectures is to centralize data stemming from several sources into a unique and conformed warehouse. In a nutshell, using a DW environment proves suitable to supply data for analytical systems within the business (Imhoff, Galemmo, and Geiger 2003), and therefore the data it holds should be generic and atomic.

3.1.3 Data Marts and Decision Support Systems

The conceptual difference between a data warehouse and a data mart is that the latter is a subset of the DW, built to answer specific business questions (Kimball 2004). Whereas a DW is specific to an enterprise, a data mart is specific to a business problem. Therefore, each business issue potentially requires its ad hoc data mart.

The structure of the data mart is not necessarily normalized. In accordance with the analytical tool, the data mart structure might follow several modeling forms (Jukic 2006). Multidimensional modeling can be used in conjunction with the star, the snowflake, or the galaxy schema for OLAP environments; whereas entity relationship modeling can be used in conjunction with normalized, flat, or hybrid schemata for data mining applications (Moody and Kortink 2000).

Data marts are used to feed decision support systems (DSS). DSS are a set of tools to conduct analyses, with the main objective of supplying information/knowledge to a decision maker on strategic issues. Two well-known examples of DSS techniques are OLAP cubes (mostly an interactive tool) and data mining algorithms (Bao and L. Zhang 2010).

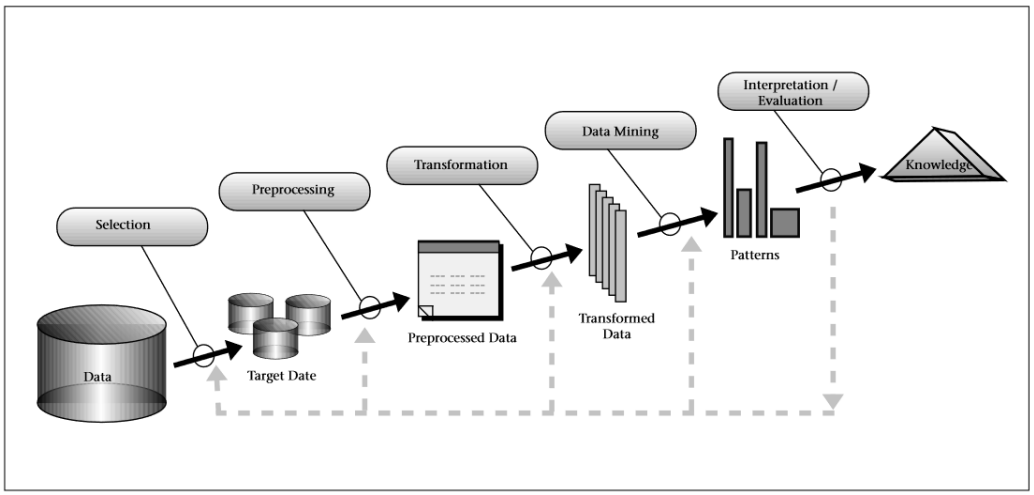
3.2 From Data to Knowledge

In this section, we describe the steps to transform data into knowledge. We start by giving an overview with the KDD process. Then, we describe some data mining techniques and introduce some concepts in fuzzy logic that can be combined with mining algorithms. And finally we present some applications of these techniques in crime analysis and forensics.

3.2.1 The *Knowledge Discovery from Data* Process

All of these architectures we have reviewed so far form an overall transformation process, each time with a different representation of information. The "data", or the element we are dealing with, is polymorphic in accordance with the level of detail and interpretation we give to it. In a simple raw log or in a database, it is considered as *data* (as no metadata describes it and no interpretation is given). When it comes to decision support systems and it is viewed by analysts, *data* turn into *information* because a context and an interpretation are given to it. Eventually, if one knows how to transform this *information* into actionable business rules, then the extraction of some useful *knowledge* is considered.

The art of applying computational algorithms to large volumes of data in order to extract knowledge is known as *data mining*. Data mining can be seen as a set of quantitative techniques where the product/output is some knowledge, or more simply put it can be seen as the activity of learning



Source: Fayyad, Piatetsky-Shapiro, and Smyth (1996)

Figure 3.1: The 5 steps of the knowledge discovery in databases (KDD) process

from data. Data mining should not be confused with the widespread term *knowledge discovery from data* (e.g., Fayyad, Piatetsky-Shapiro, and Smyth 1996; Kurgan and Musilek 2006). Knowledge discovery from data (KDD)[1]is the macro process starting from the analysis of a business problem and ending with an actionable and non-trivial pattern (Fig. 3.1). We want to emphasize that data mining methods take place in only one step of the KDD process.

3.2.2 Data Mining Techniques

Data Mining

As defined by Fayyad, Piatetsky-Shapiro, and Smyth (1996), data mining

> "involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge [...] Data mining is a step in the KDD process that consists in applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) across the data."

[1] KDD was originally defined as knowledge discovery *in databases*. As a desire to generalize the concept not only to databases, it evolved in the name of knowledge discovery *from data*.

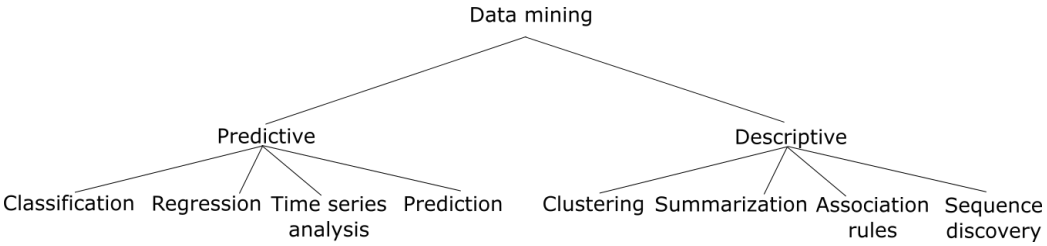


Figure 3.2: Data mining techniques classified with respect to predictive models or descriptive models.

Other definitions in the literature all share the same concept: data mining is an analytical task consisting in inferring knowledge from data.

Besides, data mining algorithms can be split into two categories, namely *supervised* and *unsupervised*, relating to their learning objective. On one hand, *classification*, *prediction*, and *estimation* are tasks considered pertaining to the former category (also known as directed data mining). Variables under investigation are split into explanatory variables and dependent variables. A model defining the relationship between these two variable types is sought. Learning the model implies knowing in advance some values for the dependent variable, which is why it is said that the dependent variable is considered as the *target*. On the other hand, *clustering*, *affinity grouping*, and *description and profiling* pertain to the latter category (also known as exploratory) (Berry and Linoff 2004).

Another way to categorize data mining techniques is in accordance with the role of their model (Dunham 2003). *Predictive models* make predictions with the use of known data. *Descriptive models* identify patterns of relationship within data (Figure 3.2).

Traditional data mining algorithms are considered as relatively mature. Definitions of these fundamentals are widespread and well documented (Witten, Frank, and Hall 2011), and many case studies give a good insight on how to use them (such as Berry and Linoff 2004; Dua and Du 2011; Westphal 2008, to name only a few). However, the impact that mining algorithms have on businesses is mostly considered as intangible, that is, difficult to measure. This latter point has often inhibited considerable investments in knowledge extraction projects within companies, where the return on investment is a decisive metric in projects assessment.

Practical aspects of data mining solutions are not to be neglected either, and technical issues can consume lots of resources. A wide-ranging review of

data mining software algorithms is conducted in Mikut and Reischl (2011). A plethora of tools are available, from completely free and open source to paying and licensed.

Domain Driven Data Mining (D^3M)

The term *domain driven data mining* (D^3M), first coined by Cao, represents a paradigm shift from *data-driven data mining*. The definition of D^3M in Cao and C. Zhang (2007) highlights the lack of intelligence and domain constraints in the traditional data mining paradigm, being only theoretical and not designed for real environments. Cao defines D^3M as

> "the set of methodologies, frameworks, approaches, techniques, tools and systems that cater for human, domain, organizational and social, and network and web factors in the environment, for the discovery and delivery of actionable knowledge."

He also suggests that domain driven data mining is an answer to "bridge the gap between academic and business data mining by providing actionable solutions, handling domain knowledge, domain factors, constraints, and in-depth patterns". A general overview of this paradigm is also provided in Cao (2008), and in Cao (2010), he explains further that D^3M has been laying the ground to other research willing to leverage this methodology. As an illustration of in-depth data mining, Cao, Schurmann, and Thang (2005) describe how to mine financial data in a stock market environment with the integration of business rules.

To support domain driven data mining, the following key components are suggested (Cao 2010):

- problem understanding and its definition should be domain-specific;
- ubiquitous intelligence is necessary through the whole KDD process;
- human roles, intelligence and involvement contribute importantly to the problem-solving and decision-support capabilities;
- data mining must deal with constrained, dynamic, distributed, heterogeneous, interactive, social, and/or networked context;
- patterns should be in-depth-ones, that is, not general patterns and using deep knowledge (such as with combined mining);
- actionable knowledge discovery is a loop-closed and iterative process

requiring refinement and acceptance by both business and technical stakeholders;

- performance evaluation should be recognized by business criterion; and
- KDD infrastructures need to be adapted.

But the most important point is to understand that domain knowledge integration starts at the data warehouse level and must be integrated within all the steps of the KDD process, especially by including domain experts during the whole process. This *ubiquitous intelligence* can be seen from several aspects (Cao 2010), such as summarized in Table 3.1.

A major difficulty about using domain knowledge arises when to decide how to represent it. Many formalisms can be used to describe domain knowledge: from ontologies to a simple additional row in a database, or by setting an acceptance threshold in a data mining algorithm in accordance with business rules. For example, an ontology-based integration of business intelligence (Cao, C. Zhang, and J. Liu 2006) explains how an element can semantically remain consistent throughout the conceptual, analytical, and physical views respectively stored in data warehouse, OLAP and data mining systems. Another piece of research (C.-A. Wu et al. 2011) draws the idea of integrated intelligent data warehouse mining based on ontologies. A more general study surveys ontology-based integration approaches, provided by Wache et al. (2001).

The challenges to face are numerous when implementing D^3M approaches, and as any other emerging research area, case studies have not yet brought empirically established effectiveness to light.

Combined Mining

Data mining activities in an enterprise-wide context imply multiple data sources. Corollary, several distinct analyses are required for these data sources, inherent to business issues. In order to better understand the business, an aggregation of these results is strongly advocated. This aggregation of patterns into sequential data mining activities is known as *combined mining*.

To illustrate this reality, let us consider a process-driven approach, e.g., guided by a crime analysis related process. This process necessarily entails analytical tasks (such as crime linkage and hot spots monitoring). These analyses are basically conducted within a closed-loop iterative environment, due to the process constraints. Patterns resulting from an analysis at the iteration i

Table 3.1: Summary of the aspects of the Ubiquitous Intelligence

Aspect	**Examples**
Data	• including data type, timing, relation, spacing, quality, etc.
Domain knowledge	• including domain knowledge, background and prior information
Human	• referring to direct or indirect involvement of humans, imaginary thinking, brainstorming, etc. • empirical knowledge • belief, request, expectation, etc.
Constraint	• including constraints from system, business process, data, knowledge, deployment, etc. • privacy • security
Organizational	• organizational factors • business process, workflow, project management • business rules, law, trust
Environmental	• surrounding business processes, workflow • linkage systems • surrounding situations and scenarios
Evaluation	• technical interestingness corresponding to a specific approach • profit, benefit, return, etc. • cost, risk, etc. • business expectation and interestingness
Deliverable and deployment	• delivery manners • embedding into business system and process

can therefore be an input of the process at the next iteration *i+1*. In more general terms, each data mining output of a task t_a can be an input of a task t_b (t_a and t_b can also be both part of the same iteration i).

A general concept of combined mining is presented in Cao, H. Zhang, et al. (2011). Furthermore, we can mention a few case studies focusing on combined mining: methods to aggregate learned association rules (Zhao et al. 2008; Plasse et al. 2007), and combined mining applied to financial reports in

Kloptchenko et al. (2004).

To sum up, we consider data mining as *combined mining* only when adhering to at least one of these conditions:

- requirements are based on multiple heterogeneous business features;
- mining is operated across multiple data sources;
- multiple data mining algorithms are combined; or
- both business and technical metrics are used to measure the interestingness of a discovered pattern.

3.2.3 Fuzzy Logic and Fuzzy Sets

Fuzzy logic, first introduced in 1965 by Zadeh based on the approach of fuzzy sets (Zadeh 1965), proposes a formalism to represent the concept of partial membership. This concept allows to assign continuous grades to an element corresponding to its membership to a class. In our particular situation, this partial membership is useful for representing uncertainties inherent to crime data. For instance, in order to classify a suspect with respect to its size, the classes *tall*, *medium* or *small* classes can be used. Indeed, an ambiguity inherent to the real-world assumes the object to be partially member of more than one class. Additionally, the same problem occurs when a hypothetical link has to be established between a suspect and a crime: it is not realistic to define this link as binomial (i.e., existent or non-existent) for inferences. This link has to be defined with a certain probability in accordance with exogenous factors. In fuzzy sets, this *membership* is described by a function lying in the real interval [0,1], representing the degree of truth for each class.

To define this type of logic more formally, we assume that fuzzy sets map an input value to its appropriate membership value, with the use of a membership function. This membership function, denoted $\mu(x)$, may be either arbitrary or any function lying within the [0,1] interval (such as a Gaussian, a piece-wise linear function, a sigmoid curve, etc.). Let X be a set of points, consisting of generic elements denoted by x ($X = \{x\}$). A *fuzzy set* is the set of pairs $A = \{(\mu(x), x)\}$, characterized by the *membership function* $\mu(x)$, describing the grade of membership of x to class A in the real interval [0,1]. Thus, as instance, when $\mu(x)$ is equal to 1, the grade of membership is the highest (meaning a full membership to the class). When this grade equals 0, it expresses the lack of membership (we can notice that these two values denote the particular case of the *crisp/boolean* logic; fuzzy logic being a superset of

the latter). This membership function can either be arbitrary or be learned with a generic neuro-fuzzy learning technique.

Conceptually, fuzzy logic is considered as a *soft computing* technique, that is, computational methods dealing with partial truth, and tolerant with impreciseness. Therefore, with respect to this context, soft computing methods can be adapted to perform reasoning with forensic data (Franke and S. Srihari 2008). To deliver more flexible approaches, fuzzy logic can be combined with data mining methods to extract crime-related knowledge.

Fuzzy Inference Systems

Fuzzy *if-then* rules are expressions in the form

$$\text{IF } A \text{ or } B \text{ THEN } C,$$

where A, B, and C are fuzzy sets. These fuzzy rules can be integrated in a reasoning system to make decisions based on imprecise modes of reasoning, similar to human thinking. A fuzzy rule consists of two parts: the antecedent (aka the premise), i.e., the term before the *THEN* clause, and the consequent. Both of these parts can have several terms, and these the terms can be combined altogether through specific operators. As an illustration, let us consider the following rule:

$$\text{IF } (\textit{risk} \text{ is } \textit{high}), \text{ THEN } (\textit{police\_presence} \text{ is } \textit{recommended}),$$

where *risk* and *police_presence* are linguistic variables; *high* and *recommended* are linguistic values characterized by membership functions.

Fuzzy inference systems provide fuzzy reasoning. These systems, when viewed as a *black box*, take crisp input(s) and generate crisp output(s). More precisely, these systems are made of five functional blocks (Jang 1993) (see Fig. 3.3):

- a *rule base* containing a number of fuzzy *if-then* rules;
- a *database* which defines the membership functions of the fuzzy sets used in the fuzzy rules;
- a *decision-making unit* which performs inference operations based on the rules;
- a *fuzzification interface* which transforms crisp inputs into degrees of match with linguistic values;

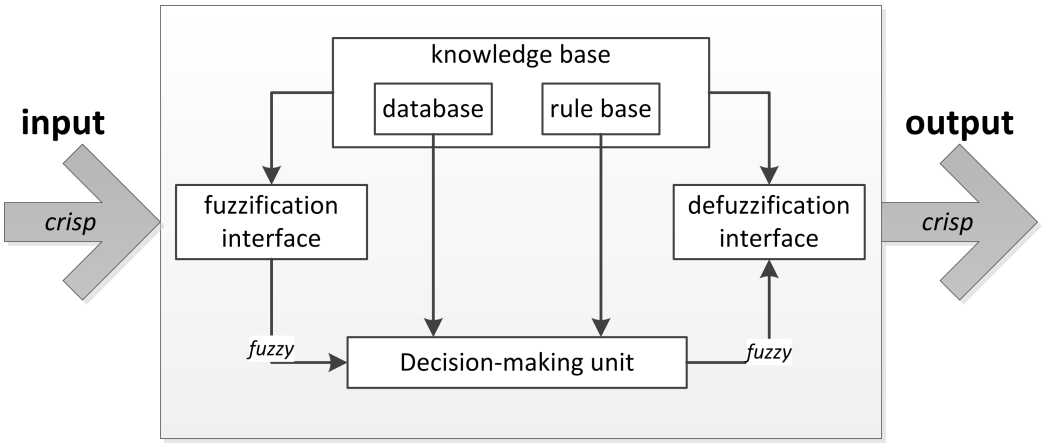


Figure 3.3: A fuzzy inference system, having both crisp inputs and outputs, consisting of 5 logical blocks.

- a *defuzzification interface* which transforms fuzzy results of the inference into crisp outputs.

Based on these features, we can now enumerate the steps of *fuzzy reasoning* (assuming that the antecedent can have multiple linguistic labels):

1. The *fuzzification*: calculation of the membership values of each linguistic label with the corresponding membership functions.
2. The *firing*: combination of the membership values of the antecedents with a T-norm operator (such as *min*).
3. The *inference*: implication from the antecedent to the consequents.
4. The *aggregation*: aggregation of the consequents according to the rules.
5. The *defuzzification*: defuzzification of the consequents to produce a crisp output.

Two widespread fuzzy inference systems can be found in the literature: the *Mamdani* type (Mamdani and Assilian 1999) and the *Sugeno-Takagi* type (Sugeno and Takagi 1985). The main difference lies in the nature of the output membership functions: in Mamdani systems they describe linguistic terms; whereas in Sugeno-Takagi systems they describe crisp values. Aside from that, as stated earlier, fuzzy inference systems can either have arbitrary membership functions, or functions learned by neural network algorithms. In the latter case, we denote these fuzzy systems as *neuro-adaptive fuzzy inference systems*.

3.2.4 Applications

Data Mining in Crime Analysis and Forensics

Data mining techniques play a major role in crime analysis and forensics. Data mining is used in crime investigation, crime detection, forensic analysis, and so on. A specific subarea of computational forensics is particularly rich of computational methods: *computer forensics*. Dealing only with computer-related evidence, it requires semi-automated methods more than ever to face the constant increase of volume of data to process. We will not concern ourselves with this specific research area, but one has to be aware that many data mining methods applied to forensics are related to computer forensics.

The spectrum of crime analysis and forensics related data mining techniques is broad. For instance, in I. Lee and Estivill-Castro (2011), a study concerning the application of data mining algorithms on massive crime data sets is presented. The famous COPLINK project provides an interesting case of multi-source data with analytical tools embedded (H. Chen, Schroeder, et al. 2003; H. Chen, W. Chung, et al. 2004). The extraction of entities from police narrative reports performed with a neural network (Chau, J. J. Xu, and H. Chen 2002). A cluster analysis for author identification is proposed in (Vel et al. 2001) and discriminant analysis (Carney and Rogers 2004). A N-gram-based technique categorizes texts from system logs (Cavnar and Trenkle 1994). Criminal profiling is applied in several different studies (Chau, J. J. Xu, and H. Chen 2002; Castellano and Sridharan 1996; H. Chen, W. Chung, et al. 2004; Hauck and H. Chen 1999; G. Wang, H. Chen, and Atabakhsh 2004; J. Xu and H. Chen 2004). An adaptive neural network is used for facial recognition (Sinha 1998). The investigation of computer logs using association rule mining is conducted in (Abraham and Vel 2002). And, finally, a specific review of forensic databases is presented in (Olivier 2009). More broadly, an extensive collection of generic methods has been gathered in the following monographs: Westphal (2008), C. Li (2009), Mena (2011), and Ray and Shenoi (2008).

Fuzzy Applications in Crime Analysis and Forensics

The applications of fuzzy logic to crime-related problems have shown considerable interest. Indeed, the inherent fuzziness of crime data requires such ad hoc systems. Fuzzy logic and its derived methods (fuzzy sets, fuzzy inference systems, fuzzy clustering, etc.) are in this sense appropriate answers.

Many traditional algorithms in artificial intelligence come in a variety of forms often using fuzzy logic. Such hybrid algorithms dedicated to finding patterns in complex data structures have been successfully applied to crime analysis and forensics. For example, fuzzy clustering is well adapted to detect/analyze crime hot spots or any other form of crime mapping (Stoffel, Han, and Cotofrei 2010; Stoffel, Cotofrei, and Han 2011; Grubesic 2006); The application of fuzzy clustering to forensic data in Liao, Tian, and T. Wang (2009), fuzzy logic and expert systems analyze computer-related crimes in the network forensics area; and fuzzy classification is used to create a text-independent automatic speaker identification in Castellano and Sridharan (1996).

Moreover, an attempt to embed some of the fuzzy sets features into criminal processes can be found in Stoffel and Cotofrei (2011). This study sets out to define a formalism —this formalism is based on BPMN (Business Process Model Notation), preserving the validity of its semantic and its syntax— to allow the representation of fuzzy attributes/constraints.

Part II

THE THREE ESSAYS

"If the facts don't fit the theory, change the facts."
—Albert Einstein (1879 – 1955)

4

An Overview of the Three Essays

This chapter gives an insight of the search process carried out and that leads to the outcomes disseminated through this thesis. Furthermore, the three main approaches developed in this dissertation are introduced.

4.1 The Search Process

A wide range of ideas has been considered and were integrated into this thesis. A part of this effort has led to several contributions: conference, seminar, and journal proceedings.

Although only the three most important contributions are thoroughly reviewed in this dissertation, I want to briefly describe every piece of research carried out during these 4 years, including shared work in which our project partners in Lausanne are the major contributors (publications in which I do not appear as first author). Presenting them in a chronological order helps to get an insight into the (re)search process.

The first paper (Albertetti and Stoffel 2012) [1]deals with the general issue of structuring crime data prior conducting analyses. Many police agencies have scarce resources and limited budget, and as a result of which very few scientific positions and even less computer science related positions are available. This lack of resources contributes to poor data warehousing solutions and even sometimes to the unawareness of this necessity. Indeed, many analytical solutions do not integrate a comprehensive view of the global information system and more specifically do not handle strategic information nor knowledge of the whole agency. To tackle this problem, a methodology to implement a data warehouse is presented. This methodology constitutes the first element of the framework.

Once having defined a reliable means of structuring crime data, general issues relating to knowledge extraction in the field of crime analysis have been promoted through various abstracts (Grossrieder et al. 2012a; Grossrieder et al. 2012b; Grossrieder, Albertetti, Stoffel, and Ribaux 2013a; Albertetti and Stoffel 2013) and presentations. Most of these contributions analyze assumptions about the predictability —or the "not completely randomness"— of crimes. Justified by criminological theories, we assume that some crimes follow certain rules and therefore, patterns or structures should be identifiable within the data.

Based on these preliminary analyses, a specific knowledge extraction method for computerized crime linkage (aka link analysis) has been described in Albertetti et al. (2013a). Crime linkage enables analysts to find series of crimes within a data set. This is a recurring issue in tactical or strategic crime analysis, whereby very few automated methods exist, although several theories

[1] First essay selected to be included in this thesis (Chapter 5)

suggest the presence of repeated commonalities or "signatures" through specific crimes.

As an extension to this study, a more detailed paper (Albertetti et al. 2013b)[2]further explains the idea by providing a detailed methodology. Moreover, a relevant analysis of underlying criminological assumptions is also carried out.

As a matter of fact, devising and implementing computerized methods requires some deep knowledge and understanding of the domain specificities, especially in crime analysis where inferences are performed on partial and uncertain data. To assess the overall applicability of data mining techniques in crime analysis, an extended study has been proposed in Grossrieder, Albertetti, Stoffel, and Ribaux (2013b). The study puts an emphasis on the requirements for producing suitable intelligence; for that one has to think over the characteristics of both crime analysis and data mining. These requirements led to the design of the Computational Forensic Criminology framework (Grossrieder, Albertetti, Stoffel, and Ribaux 2015). This framework provides the context in which computational methods should be applied to solve crime analysis problems.

Exploring a second knowledge extraction method based on our previous results, it appeared that data mining techniques are particularly suitable to support follow-up and detection systems in crime analysis. As a consequence, we decided to direct the focus towards the detection of changes in crime trends. Indeed, from both operational and strategic sides, the detection of changes in trends turns out to be particularly useful. It gives crime analysts the opportunity to better link activities from the same offender (or the same group). It also enables them to more easily understand and discover hot spots and patterns, to detect repeat victimization, and more generally to better understand and detect significant changes in crime trends. As a first answer to this problem, we explored the *change points detection* literature. We tested some existing methods against our data set. In accordance with the outcomes, a polynomial basis for modeling time series has been put forward. These preliminary results have been disseminated through Grossrieder, Albertetti, Stoffel, and Ribaux (2014b) and Grossrieder, Albertetti, Stoffel, and Ribaux (2014a).

Finally, the last piece of research further describes change points detection in Albertetti, Grossrieder, et al. (2016)[3]. The proposed method detects changes within crime-related time series in two steps. First a segmentation is performed to the time series, then these segments are queried using a flexible

[2] Second essay selected to be included in this thesis (Chapter 6)

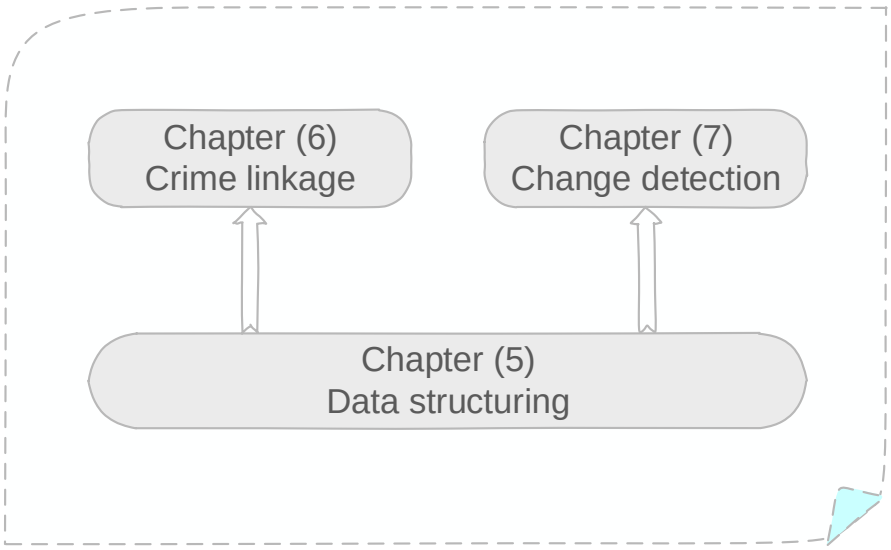


Figure 4.1: Overview of the three essays. The first essay (data structuring) acts as a basis supporting the two others (crime linkage and change detection).

approach based on fuzzy sets.

4.2 Presentation of the Three Essays

In this section the three selected essays are presented. The first essay (Chap. 5) deals with organizing and structuring crime data suitably for conducting data mining algorithms, which is one of the first concerns of this dissertation. Then, the second essay (Chap. 6) and the third essay (Chap. 7) are "instances" of applied knowledge extraction methods to crime analysis, considered as products and examples of the proposed framework, respectively addressing the issue of crime linkage and of change detection. Both methods rely on the outcome of the first essay, that is, first data are transformed with the structuring proposed method and only then this product is utilized by the mining algorithms (Fig. 4.1).

[3] Third essay selected to be included in this thesis (Chapter 7)

4.2.1 Structuring Crime Data: from Police Reports to Data Marts

The first essay (Albertetti and Stoffel 2012) is about preparing crime data with the purpose of conducting mining analyses. As a general observation, many mining methods delivered to crime analysts simply "omit" to mention that a preprocessing environment should prepare crime data with respect to its inherent nature before any kind of analytical method can be applied. With this essay, we provide a first step of the framework for structuring crime data with concepts grounded in the data warehousing theory. The proposed methodology details a 5-step process to structure crime data stemming from police reports. As a result, data marts specifically designed in accordance with the requirements ensure that analytical techniques can be correctly applied and fully exploit the information imbued within the data. For instance, a specific data mart has to be created to apply data mining classification techniques, another specific data mart to explore data with a dashboard, a histogram for time series analysis, and so on. In particular, classification data mining techniques and fuzzy methods to deal with uncertain data have been considered.

One of the main outcomes of the methodology is the ability to integrate intelligence or metadata in the overall process of analysis. Because this intelligence is added in the data warehouse step, it can be reused through any kind of analysis. Furthermore, this augmented situation gives a context of the data and therefore specifies how to handle, to some extent, both uncertain data and unclear objectives.

4.2.2 A Method for Crime Linkage

The second essay (Albertetti et al. 2013b) describes a computerized crime linkage method. Consisting in determining if a set of offenses has been committed by the same offender, crime linkage enables to find series of crimes within a data set. In the proposed approach, we first define a similarity coefficient between two crimes based on multi-criteria decision making and fuzzy sets. Contextual, physical and behavioral information is considered, for both high-volume crimes and serious crimes. Second, with some data stemming from the State Police of the Canton de Vaud through our partnership with the University of Lausanne, we present a proof of concept and evaluate the results.

Furthermore, some underlying criminological assumptions are presented. These assumptions are critical key success factors to an effective implementation of any crime linkage system. Besides, some guidelines to evaluate the outputs of such system are proposed.

4.2.3 A Method for Detecting Changes in Crime Trends

The third essay (Albertetti, Grossrieder, et al. 2016) proposes a change points detection method for crime analysis. The intuition of the method is to be able to detect and query change points among any time series with a flexible and understandable approach. By enabling linguistic terms to express query patterns, the search of trends becomes easier, especially for domain expert not familiar with data mining techniques.

The proposed method entails three steps. (a) A time series is represented by the means of a polynomial model. Some basic orthogonal functions are expanded; the parameters of this expansion are found to be optimal in a least-square sense. By choosing the particular discrete Chebyshev polynomials as basic orthogonal functions, the parameters of the expansion find an interpretation as estimators of average, slope, curvature, etc. This characteristic provides the advantage to express more easily a particular change in a trend in terms of linguistic variables. (b) The time series is split into multiple subsequences. This window segmentation has the advantage to be dynamic, i.e., the segments will be defined in accordance with the shape of the time series. To achieve this, some thresholds are defined in terms of shape estimators (the parameters of the expansion). And (c), a fuzzy inference system is used in addition to find the appropriate subsequences relating to the queries stated using linguistic terms.

"Good wine only comes from good grapes."
—Wine grower proverb

5

Structuring Crime Data: from Police Reports to Data Marts

Many crime analyses are conducted with computational methods directly derived from the data mining field. Most of these methods are dedicated to a particular task, therefore crime analysts use a number of them. Because of this, the need to provide a persistent basis for the data to analyze becomes more important than ever, especially with the specificities of crime data. However, very few of these methods integrate this perspective. In this paper we present a methodology to structure police report data for crime analysis. The proposed artifact is mainly about applying data warehousing concepts to crime data in a crime analysis perspective, with the objective of including intelligence that can handle uncertain data and unclear objectives. Moreover, a proof of concept is carried out with real forensic data to illustrate and evaluate our methodology. These experiments highlight the need for such framework for crime analysis.

This essay can be found as "From Police Reports to Datamarts: Towards a Crime Analysis Framework," Albertetti and Stoffel (2012).

5.1 Introduction

Over the last decades digital treatment of data and information related to criminal activities has gained drastically in importance. Usually, this information is stored in police logs or police reports (hereafter referred to as "logs" for the sake of brevity). Such information may serve as a basis for crime investigation or may provide important pieces of evidence in prosecution perspectives (Hess, Orthmann, and Cho 2010). One specific type of forensic data we will consider in the remainder of this chapter are documents dealing with criminal activities as well as police activities, containing mainly crime incidents, arrests, calls for service, and accidents. Generally focused on pragmatic and operational needs, the structure of these logs is not designed for decision support systems (DSS) (Inmon 2002; Kimball 1997) —data mining, on-line analytical processing (OLAP), reporting, dashboard, etc.— and hence not for complex crime analysis. Despite this fact, in the case where logs are directly stored in a database, such forensic sources may still be directly used as an analytical tool by some analysts. To have such "turn key" solutions may have numerous advantages such as decreasing costs or reducing development time to name only two, but none of them have theoretical justifications.

In this chapter, we propose a methodology to fill this gap. It leverages the inherent information in forensic data by specifying in a first step a wider analytic framework. The proposed artifact enables this feature by applying well-established methods from the field of business intelligence to the computational forensics discipline. More particularly, the concepts we propose belong to the data warehousing area: corporate information factory, multidimensional modeling, and its underlying data marts, are the main ones. The proposed methodology, consisting of a process and some key success factors, is about designing data models. Once these models are implemented, the crime analysis framework is ready to be used by a crime analyst (i.e., the use of specifically required DSS will be supported/enabled by an underlying ad hoc architecture for data based on the logs).

Another advantage resulting from the application of data warehousing techniques is to be consistent with intelligence-led policing models (e.g., Ratcliffe 2008): actually, the latter paradigm highlights the lack of proactive strategies in traditional policing models and advocates a partnership among all police departments involved. These disadvantages are partially mitigated by breaking down traditional organizational stovepipes with a unified data structure – the data warehouse – and converging towards the production of forensic intelligence through advanced analyses (e.g., the extraction of knowledge by

using, for example, data mining techniques).

Assuming the logs to be stored in a transactional/operational database, we identified five main iterative tasks defining the overall approach. Moreover, meta-requirements are defined as key success factors to be followed during the process steps. These key success factors act as main thread and are transversal guidelines to potentially carry out the methodology successfully.

These five main tasks can be summarized as: (1) targeting the required data and identifying forensic challenges, (2) designing a business data model, (3) gathering the required transactional data into a generic data warehouse (DW), (4) building dedicated data marts for each required analysis (e.g., for DSS) to find answers to crime issues, and (5) implementing and testing the framework. The approach attempts to be generic by dealing with recurrent crime analyses issues. Furthermore, some examples are provided to illustrate specific issues stemming specifically from forensic data (e.g., the time an event occurs requires an adequate definition in the DW, by taking into account daylight saving time in order to better analyze burglary phenomena).

Beside this, these following questions are addressed:

- Do analytical biases decrease by having an ad hoc data structure?
- How to conduct an analysis based on multiple data sources?
- How to organize logs to get normalized and consistent data?

The remainder of this chapter is structured as follows: Sect. 5.2 presents related research and highlights the need of the proposed artifact, Sect. 5.3 lays the groundwork by introducing briefly some data warehousing and data mart concepts. Then, the proposed methodology is described in Sect. 5.4 in a way such to be as generic as possible in respect to forensic issues and illustrated in Sect. 5.5. Sec. 5.6 evaluates the methodology through a case study based on real forensic data (police reports). To conclude, we postulate that an efficient data warehouse and its subsequent data marts contribute to a better crime analysis framework and discuss the idea of a potentially wider analytical framework including this artifact as a starting point.

5.2 State of the Art

Computational forensics is about applying *computational* methods from several disciplines in the forensic domain. In the sense of Franke and S. Srihari (2008),

these methods support forensic sciences in three ways: (a) they provide tools to overcome limitations of human cognitive ability, (b) a large volume of data is potentially usable for analyses and is not anymore limited to the human mind, and (c) human expert knowledge may be numerically represented to teach inference machines.

Many applications of computational forensics have been approached following these three basic ideas. From the evolution of automated fingerprint identification systems (Maltoni et al. 2009) or methods for hot spot crime detection (Grubesic 2006), to a fuzzy extended BPMN (business process model notation) for modeling crime analysis processes (Stoffel and Cotofrei 2011), by an analysis of crime measurement and statistics with the national incident-based reporting system (aka the NIBRS) (Maxfield 1999), to name only a few.

Other pieces of research emphasize on the role of data and their importance to be well organized. In Ribaux, Walsh, and Margot (2006) it is reported that forensic case data is poorly integrated into crime investigation and crime analysis. They advocate the use of a framework for putting in advance forensic case data, in order to go beyond the single production of this latter type of data for court evidence. In Ferguson (1997), we can see how companies were able to answer more quickly to business issues when using data warehousing. Mikut and Reischl (2011) present an overview of data mining tools, and emphasize on the need of an appropriate data warehouse. A general framework for crime data mining has been briefly approached (H. Chen, W. Chung, et al. 2004). Finally, an analytical methodology is proposed by Westphal (2008) to apply data mining focused on a criminal detection perspective.

However, none of these studies explicitly combine computational forensics with methods to structure data in an analysis perspective. For instance, crime analysis is very seldom introduced by specifying an underlying structure that forensic data require. This is the gap we want to fill with this chapter.

5.3 Data Structures and its Systems

In this section, a brief definition and an overview of the covered subjects is presented. As a starting point, crime analysis can be defined as

> "the systematic study of crime and disorder problems as well as other police-related issues (including socio-demographic, spatial, and temporal factors) to assist the police in criminal apprehension,

> crime and disorder reduction, crime prevention, and evaluation". (Boba 2009)

In this chapter, we focus on a specific aspect of that definition, i.e., crime analysis in conjunction with the use of information technologies and computational methods. The main objective is to understand that the data to analyze might be structured in several ways.

5.3.1 Transactional/Operational Systems

A transactional/operational system, dealing with day-to-day transactions, usually uses data stored in a database supported by a database management system. This kind of environment is known as *on-line transaction processing*. Not really designed for long term analyses or strategic purposes, data are considered to be structured and normalized (usually at the third normal form or Boyce-Codd normal form) to avoid redundancy and to preserve consistency. Conceptual entity-relationship (E-R) or relational models (contrasted with multidimensional models) are techniques used to draw these systems.

5.3.2 Data Warehousing Systems

Whereas operational systems support business processes, data warehouse systems support evaluation/analysis processes. A widely accepted definition of a data warehouse is "a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making" (Inmon 2002). We contrast this statement by arguing that a DW does not directly integrates decision making and is not subject-oriented: these features are transferred to another level, the data marts. So the DW is not directly serving or supporting a DSS, but the data mart is. Actually, the general purpose of such data architecture is to centralize data stemming from several sources into a unique and conformed warehouse. In a nutshell, using a DW environment has proved to supply data for any form of analytical system within the business (Imhoff, Galemmo, and Geiger 2003), and therefore the data should be generic and atomic.

Police logs are assumed to pertain to an operational perspective, i.e., focused on day-to-day purposes and supporting short term objectives. On-line transaction processing systems deal with such data. For example, an event (which could be an incident in a police event log context) is characterized by its date, the persons concerned, the area where it happened, etc. This

event, once reported, is recorded in a police event log, physically stored in an on-line analytical processing system. Policemen may access to it, modify the logs, add some comments and perhaps use it in a court as a piece of evidence. Whereas when it turns to crime analysis or crime investigation, another approach is followed. Crime series, links between people, or any other information belonging to the investigative process are subject to different requirements; a deeper data analysis is conducted and hence a "basic" operational framework is everything but an efficient approach. Data analysis using time series, statistical studies or any other analysis implying the notion of aggregated information (such as the aggregation of months by season) assumes using a DW architecture.

5.3.3 Data Marts and Decision Support Systems

In the previous part, we explained the role of a data warehouse, and emphasized on the fact that a DW is designed to be unique and generic. The conceptual difference with data marts is that they are a subset of a DW, built to answer specific business questions (Kimball 2004). Whereas a DW is specific to an enterprise, a data mart is specific to a business problem. Therefore, each business issue potentially needs an ad hoc data mart.

The structure of a data mart is not necessarily in a normalized form; in accordance with the required structure of the analytical tool, it might be in any form (Jukic 2006). Multidimensional modeling may produce star, snowflake, or galaxy schemata in an OLAP environment whereas entity relationship modeling may produce normalized, flat, or hybrid schemata for data mining applications (Moody and Kortink 2000).

Decision support systems are a set of tools fed by data marts to conduct analyses. Some examples of DSS techniques are OLAP cubes or data mining algorithms (Bao and L. Zhang 2010).

Basically, the "data" we are dealing with are polymorphic according to the level of detail and interpretation we give to it: in a simple raw log or in a database, it is considered as *data* (as no metadata describes it and no interpretation is given); when it comes to the DSS and is viewed by analysts, *data* turns into *information* because a context is given to it; eventually, if we know how to use this *information* into actionable business rules, then we consider some useful *knowledge* may be extracted.

5.4 The Proposed Methodology

The essence of the methodology we propose here pertains to the art of data warehousing design; however, the specificities resulting from the application to the computational forensics domain are taken into account and have a considerable impact on the proposed artifact. It is not intended to explicitly define technical details or implementation issues in this chapter, however, some concepts will be illustrated in the next section (Sect. 5.5).

Conceptually, the methodology is a recipe helping in switching from a data acquisition environment to a data delivery environment (i.e., a framework helping in converting operational data into business intelligence or knowledge).

As mentioned before, police event logs are assumed to be stored in databases. These databases use as the finest grain of data a criminal *incident*, meaning each incident/event occurring turns into a new line (i.e., a *transaction*) within the database. These incidents often share the same structure among different police logs and therefore can be usually summarized by (Boba 2009):

- an incident number (acting as a transaction identifier),
- the date of the report,
- the location of the crime,
- the date of the incident,
- the method of the crime (aka *modus operandi*), and
- sometimes a description.

5.4.1 The 5-Step Iterative Process

The 5-step iterative process we propose uses an operational data structure as a starting point and ends by the creation of data marts. Based on a process used to design a data warehouse to represent the enterprise perspective (Imhoff, Galemmo, and Geiger 2003) according to the corporate information factory (CIF), it has been simplified and adapted to a shorter iterative cycle. This process is iterative and therefore the cycle has to be performed several times to converge towards the most appropriate solution. The resulting required steps we identified are depicted in Fig. 5.1.

The first task is to identify the data needed to conduct crime analyses. As part of an iterative process, it is not sought to set exact boundaries for the

first time. A set of business questions needs to be raised by crime analysts. Next, we must identify the data sources required to be able to answer to these questions. In our case, the only source might be the logs or a subset of the logs. A clear idea of the main elements or topics to be included in the DW is important. A subject area model depicting the main entities and their links is the principal output of this step. This model serves as a reference to go further and as a communication basis among stakeholders.

Then, the second step refers to a conceptual data model. This detailed model, representing at a more specific level of detail the relationships and the attributes of the subject area model, is helpful to understand the business. This model is called the *business data model* and is generally using the entity-relationship (E-R) modeling technique. If information stemming from logs is already stored in a database, then the structure is already defined through the system and the business data model will look the same except technical details. Otherwise, a model has to be designed to lay the ground for the next steps. The main advantage is that this model helps people to envisioning how these different parts fit together. This model is not meant to be directly used (neither in a transactional nor in an analytical environment), it is purely an intermediate step.

The third task consists in designing the data warehouse. Designing a DW is the most challenging part of this methodology, and could be the subject of an entire book. Many conceptual choices have to be made, and these decisions are mainly part of the information systems domain knowledge. Nevertheless, as we are dealing with a specific case —an ad hoc DW about police reports for the purpose of crime analysis—, some "shortcuts" can be taken and consequently the level of abstraction decreases.

The purpose of building such an intermediary layer is to perform the tedious processes of extracting, transforming and loading (ETL) data from the transactional environment into the warehouse only once, and to lay the ground for the further data marts. According to Inmon (2002), the DW is the foundation of all the DSS processing. Its role is to act as an intermediary layer between the operational and the analytical environment. As stated in Section 5.3, a DW is normalized and the design technique we recommend is the E-R modeling technique.

The fourth step entails the creation of data marts. Whereas the purpose of a DW is to be generic and normalized, data marts are shaped by requirements and have to be suitable for the DSS, which often implies a multidimensional modeling approach (Inmon 2002). Actually, a data mart is specific to a subject, to a business problem, and to an analytical tool. The recommended

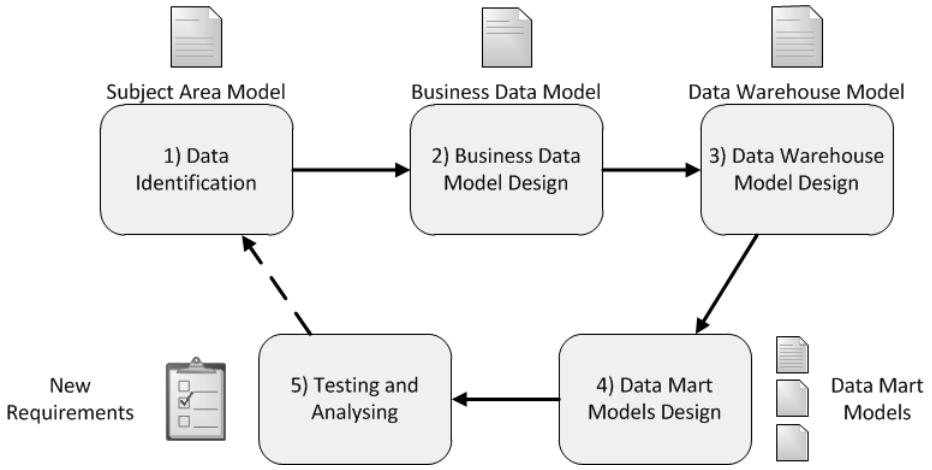


Figure 5.1: The 5-step iterative process of the proposed methodology

methodology is to design a data mart for each of the possible combinations. Data marts can be viewed as replicated subsets of a data warehouse, and the data within are not meant to be updated (the only atomic and consistent *source* being the DW).

The last step is all about implementing, testing, and analyzing what is missing for the next process iteration.

The implementation is mainly about extracting, transforming and loading (ETL) data from a system to another. The ETL technique will not be presented here, but we would like to underline that one should not underestimate this step (in a matter of time).

Testing the system means checking the integrity of the data (Are the data of the new system consistent and coherent with the data of the source system?), checking the interoperability with the DSS, and measuring its effectiveness (i.e., the capacity of the system to answer to business questions mentioned by analysts according, to the time and the resources needed). From these tasks, considering the system is never perfect, should emerge new requirements, being future inputs of the next iterations of the process.

We want to emphasize that users of such systems are crime analysts first and foremost, and technicians second. Once the ETL project is defined, pushing a single button should be enough to activate the necessary mechanism to update the data flow through all the steps.

Table 5.1: Key success factors for designing and implementing a data warehouse

Key Success Factor	**Steps**	**Effects**	**Based on**
Design the DW with a generic approach	1,2,3	Reduces the development time for a marginal DSS, decouples DW from DSS and technologies	Imhoff, Galemmo, and Geiger (2003)
Manage the data quality upstream	1,2,3,4	Avoid redundant work, preserves data consistency	Inmon (2002)
Guide conceptual choices with forensic answers	1,2,3,4	Business driven approach, more realistic	Kimball (2004)
Use an iterative approach	All	Lighter development cycle, easier to integrate new requirements	Imhoff, Galemmo, and Geiger (2003) and Inmon (2002)

5.4.2 Key Success Factors

Key success factors are a list of requirements that need to be fulfilled to considerably increase the rate of success. However, there is no absolute guarantee that whether the key success factors are respected the methodology will be systematically successful.

Each of the specified factors (detailed in Table 5.1) includes the concerned steps of the process, the positive potential effects on the methodology and their theoretical groundings. They have to be seen as general guidelines to follow during the implementation of the process, or as meta-information about the process.

5.5 Proof of Concept

The main purpose of this section is twofold: (1) to demonstrate the capability of the proposed methodology to enhance crime analyses through the use of DSS and to show the benefits of building a DW as an intermediary layer

(i.e., reducing marginal costs/time of building specific data marts or ad hoc environments for each specific crime analysis); and (2) to illustrate the methodology through a case study with data stemming from a real environment.

As a starting point, we will describe the data used in the remainder of this section. In partnership with the Police de sûreté du Canton de Vaud (Switzerland), we gathered a subset of data stemming from their police event logs. Mainly focusing on burglaries and being event-centric, these data have several purposes within the organization: keeping a trace of all events for court evidence, serving as a directory for simple exploration, feeding an operational tool used by crime analysts in order to support police operations and to better understand crime series, etc. The input we used to anchor the first step of the proposed methodology is composed of several files encoded in the comma separated values (CSV) format. This is generally the effect of exporting data from other existing tools, such as the one used to store the logs.

These data, because gathered in a police reporting context, are not suitable for advanced analyses. For example, the date of the event is not detailed (the day of the week is not directly given; it is not stated if the date is a public holiday; etc.) and is not accurate (the moment of an event is not a unique date: it is an interval given by the victim —usually corresponding to the interval from when the victim leaves one's house to when coming back). Moreover, the time of the event has no context, in the sense that neither the daylight saving time nor the time zone is mentioned. Another example is the GPS coordinates of the event: the location format uses Swiss Cartesian coordinates (a specific X-Y-Z system). Depending on the import function of the geographic information system (GIS), the data might need to be converted. More generally, these problems can be summarized by this statement: *there is no metadata giving a context to interpret/analyze these events with computational methods.*

Another fact to notice is that each event may have several modus operandis (and vice-versa), therefore a many-to-many relationship is required. This kind of relationships (usually occurring for *categories* in general) may be a problem when it comes to modeling the DW. If the DW were not generic, then we would denormalize these categories into an array of fixed sized (i.e., if the maximal cardinality of an event is of N categories, then an array of N categories is denormalized in the event table). This choice is not appropriate because it limits the use of other potentially useful DSSs (certain requiring normalized data).

In accordance with the first step of the methodology and in order to identify

the data required for crime analysis, some business questions have to be raised by the end users conducting analyses in the future DSS. These questions might prove essential to delineate the scope and to emphasize the added value of a crime analysis framework. Here are a few examples of questions that could be considered by crime analysts focusing on burglary issues: "Can we group some events together with specific burglary phenomena in order to better understand them? What kinds of crime methods often occur together? Are some crime methods following a predictable trend? What is the seasonal trend of burglaries using a window to enter a house as modus operandi? How many events during summer imply many offenders and last more than one hour?" It is assumed that these questions require data mining techniques and perhaps an OLAP system, due to the inherent difficulty to provide answers using simple SQL requests. A subject area diagram describing the main entities may be drawn from these questions. Entities such as event, modus operandi, and location of the crime are obviously needed.

Then, the second step of the process is to design the business data model for describing how these entities are constituted and how they interact. This business data model, normalized and using a relational modeling technique, must handle all needed attributes for the analysis. For example, the location of a crime will be defined in a way to handle all the attributes present in the log, but in a normalized way (e.g., if the policeman reported the modus operandi of the crime as free text, a normalized entity regrouping these values has to be included in the model).

To illustrate some issues of the DW design, let us consider a date describing the moment an event took place. In the previous steps, the structure of the data was probably a single field stating the day, the month and the year with a separator (e.g., *03-06-11*), which was totally sufficient. Whereas in the DW, another structure is needed in accordance with our business questions: if we want to conduct an analysis based on the seasons or to forecast a trend related to weekdays or public holidays, a detailed *calendar* has to be implemented to make the navigation across temporal aspect easier. This can be translated by the creation of entities including this metadata (see Fig. 5.2: entities *Dates, Days, Months and Years* have been created to represent the concept of a calendar; the derived field *duration_h* has been added to easily retrieve the duration of an event; etc.). The creation of metadata can be leveraged in the same way to consider other aspects such as the daylight time saving, holidays, location, or any other hierarchy.

The fourth step is about designing data marts. In accordance with the numerous DSS requirements, their respective ad hoc data marts will be

designed. To analyze our data, we decided to use the following DSSs: one for a data mining system and another for an OLAP system. While the mining mart needs a flat structure (a single denormalized table) using an E-R model, the OLAP mart needs a star schema (modeled with the multidimensional technique).

Without the DW layer, we can imagine the problems we would face when implementing these two data marts. All the issues about the date, the coordinate systems, and the categories would have been done at least twice. The complexity of handling such systems with no intermediate DW is not bearable when even more data marts are considered.

The mining mart (see Fig. 5.3) is designed in accordance with the specifications of the data mining algorithms included in the data mining software chosen (*Weka* in our case). As most of the work had been done during the design of the DW, only little adaptation is needed: one of the most difficult task is to denormalize the N-N relationship between the table *Events* and the table *Modus*. The same holds for the OLAP mart: the calendar is already designed and only a few changes (such as denormalization) are required.

The last step entails the implementation and the testing of the framework. The implementation part extracts, transforms, and loads (with the use of ETL techniques) to move gradually the data from the raw logs to the DW and finally to the data marts supporting the DSSs. Unifying and cleaning the data is very important to guarantee that the integrity and the conformity of the data is ensured all along the processes (e.g., the event *evt306-fis* which occurred in Lausanne at 9pm on Friday 03-09-11 described in the log needs to be recognizable as such throughout all the steps). Lastly, the testing part has to be done by crime analysts. In our case, some data mining techniques were used to answer some business questions. These tests have defined new needs as well as new business questions. As such, they fulfill the requirements of an iterative process.

All the software used within this part is open-source software. The database management system is based on a MySQL server (from Oracle), the ETL part is using Spoon (being a subpart of Kettle PDI from Pentaho), the data mining DSS is Weka (from Pentaho) and the diagrams were created with MySQL Workbench (from Oracle).

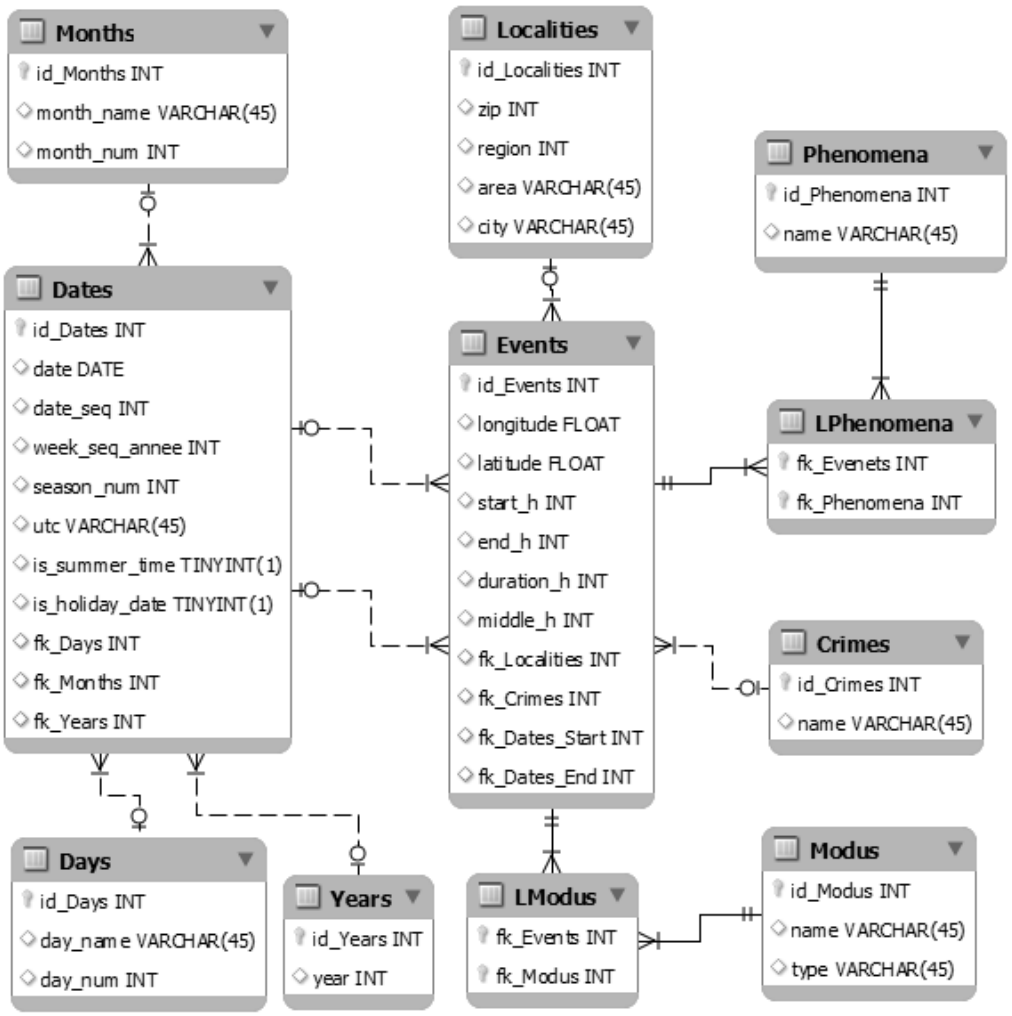
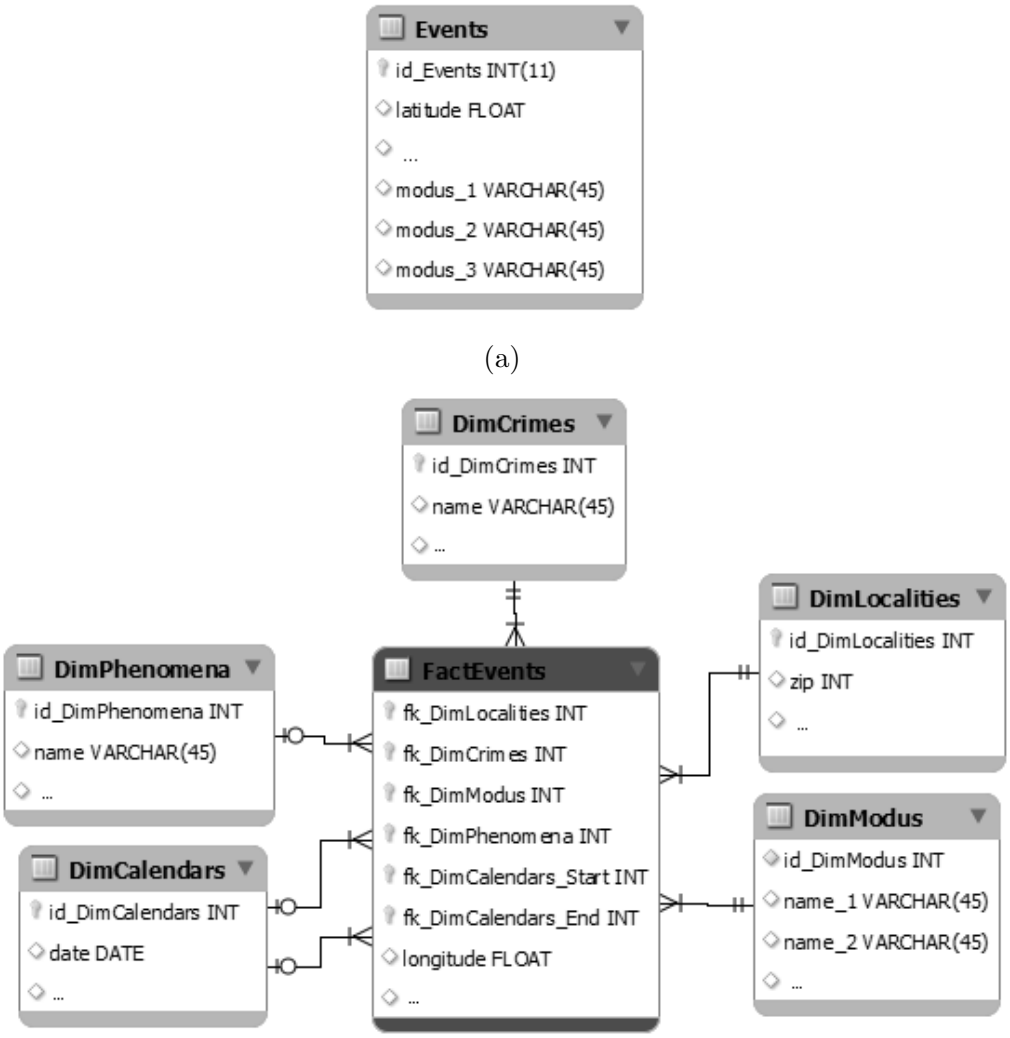


Figure 5.2: A data warehouse diagram representing all required entities for burglaries analyses in our case study.



(a)

(b)

Figure 5.3: Excerpts of data mart diagrams resulting from the latter DW: (a) a data mart for a data mining DSS (Weka) and (b) an OLAP mart for the fact *Events*.

5.6 Evaluation

In this part, we aim to evaluate the usefulness of the proposed methodology. To do so, we suggest to use the following criterion: What is the difference between a scenario when an analysis is conducted without implementing an advanced structure (i.e., analyses conducted from raw logs) and a scenario when the analysis is conducted with the use of our artifact?

To conduct the evaluation, we will consider this basic scenario: "An analyst investigating burglaries wants to understand the significance of a day to be a public holiday (for the year 2011)". Hence he probably needs to calculate (a) the probability that a burglary was committed during a public holiday and (b) the probability that a burglary was committed when that day is not a public holiday.

The latter calculation is not difficult, and only three measures are needed: the total number of burglaries (for the considered year), the number of burglaries that happened during public holidays, and the number of burglaries that happened during working days. Considering finding these numbers with data stemming from logs (i.e., without the use of the framework), a few steps will be needed (as extracting the day of the date and looking up the date in a calendar including public holidays). Whereas with the implemented framework, the latter work has been done already and can be leveraged in all further similar analyses: we just need to select the events by filtering them in accordance with the column *is-holiday-date* and to count them.

Of course, in both cases we may be able to reach the objective, but as the number of business questions increases and as the computations become more complex, the first scenario is de facto inadequate and too laborious compared to the simplicity of the questions the crime analyst may raise (i.e., not much efficient). The second scenario facilitates the task of the end user by delegating these technical tasks to another person (the person designing and implementing the data warehouse) and by avoiding redundant work. Another important fact has to be noticed, namely that the risk to introduce errors into an analysis process is reduced, as most of the work is zoned directly by the DW, always following the same procedure.

5.7 Conclusions

The main purpose of this chapter is to propose an artifact describing how to lay the groundwork for crime analysis by specifying a way to manage forensic data.

By using an intermediate layer to store the data in a structured way (i.e., a data warehouse), we illustrated how crime analysts avoid redundant work; reduce the risk to introduce cognitive biases; move from a "stovepipe" view to a transversal view helping in inter units communication and therefore converging towards an intelligence-led policing approach; and decrease the marginal cost and the time of answering new business questions.

We also evaluated the usefulness and the applicability of the proposed methodology by comparing different scenarios. The methodology was successfully implemented, and the main finding is to make much easier and less tedious the running of data analyses in order to answer business questions.

Being a first step towards a wider crime analysis framework by structuring forensic data serving as a basis for computational methods, we obviously encourage additional work or discussions.

Acknowledgements

The authors would like to thank the Police de sûreté du Canton de Vaud and its chief Alexandre Girod for their cooperation and the data provided. We are also grateful to the Swiss National Science Foundation (SNSF) for financial support.

"All traditional logic habitually assumes that precise symbols are being employed. It is therefore not applicable to this terrestrial life but only to an imagined celestial existence"

—Bertrand Russell (1872 – 1970)

6

A Method for Crime Linkage

Grouping events having similarities has always been interesting for analysts. Actually, when a label is put on top of a set of events to denote they share common properties, the automation and the capability to conduct reasoning with this set drastically increase. This is particularly true when considering criminal events for crime analysts; conjunction, interpretation and explanation can be key success factors to apprehend criminals. In this chapter, we present the CriLiM methodology for investigating both serious and high-volume crime. Our artifact consists in implementing a tailored computerized crime linkage system based on a fuzzy MCDM approach in order to combine situational, behavioral, and forensic information. As a proof of concept, series in burglaries are examined from real data and compared to expert results.

This essay can be found as "The CriLiM Methodology: Crime Linkage with a Fuzzy MCDM Approach," Albertetti et al. (2013b).

6.1 Introduction

Crime linkage is quite a particular challenge. As stated by Grubin et al., crime linkage consists in determining if a set of events has been committed by the same offender (Grubin et al. 2001). Basically, two tracks provide an answer to this problem: behavioral and situational similarities. The first aspect relies on the behavior of the offender, described by particular methods of crime (aka *modus operandi*). The second is based on spatiotemporal similarities: analyses are conducted in accordance with crime mapping theories. Furthermore, a third aspect should also be considered but is often omitted: forensic cased data (such as DNA, shoemarks or earmarks). One of the main reasons of this omission is that, depending on the organizations, the availability of forensic data within the criminal investigation process is limited. Practically, it also assumes that the offender physically leaves a trace of a satisfactory quality in more than one crime scene. Nevertheless, forensic information is much more reliable and is becoming more integrated into intelligence databases (Rossy et al. 2013).

As several studies primarily focus on behavioral linkage analysis and on situational analysis (e.g., Grubin et al. 2001; Santtila et al. 2008; Hazelwood and J. I. Warren 2004; Melnyk et al. 2011), this current study attempts to combine behavioral, situational, and forensic information. Covering these three aspects is a way to combine techniques for both serious and high-volume crimes analysis. To do so, several mathematical methods that effectively compute these similarities can be put altogether. Our approach is based on a fuzzy multi-criteria decision making (MCDM) method in order to deal with that issue. Indeed, in a crime analysis context, fuzziness permits modeling experts' experience and handle vagueness whereas MCDM is useful to evaluate and combine similarities stemming from multiple criteria.

The remainder of this chapter is structured as follows: Section 6.2 introduces the key concepts in both fuzzy MCDM and crime linkage. A review of similar research is also presented. In Section 6.3, we detail the proposed methodology with the objective of implementing a computerized crime linkage system. A proof of concept is described in Section 6.4 with real data about serial burglaries. To sum up, Section 6.5 concludes on the perspectives of the artifact.

6.2 Literature Review

In this section, multi-criteria decision making concepts are introduced and a comparison of the underlying techniques is undertaken. Then, fuzzy sets are presented with the focus on translating experts' informal reasoning knowledge into membership functions. Third, we review a set of related work focusing on combining these methods. To finish, a forensic framework is set up with the purpose of revealing the big picture of the problematics we are confronting.

6.2.1 Multi-criteria Decision Making and Fuzziness

Multi-criteria decision making is a vast research area. However, the main concept is simple: helping decision-makers in choosing, ranking, or sorting alternatives according to a set of criteria. In the literature, two basic approaches are frequently used: multiple objective decision making and multiple attribute decision making (MADM). Our interest lies in this latter: MADM.

In an MADM problem, the *decision maker* wants to select a particular subset from a set of *alternatives*. These alternatives are described by some *criteria* (at least two), and each criterion can have a different importance (the *weight*) in regard to the other criteria. These weights usually reflect the subjectivity of the decision maker. From that point of view, three challenges can be thought of: (a) finding the best alternative (*choice* problem); (b) ranking all the alternatives in decreasing order (*ranking* problem); or (c), assigning the alternatives to predefined ordered categories (*sorting* problem).

Plenty of MADM methods can be used to tackle these problems, such as maximin, maximax, AHP, TOPSIS, or multi-attribute utility theory (MAUT). MAUT is an interesting approach and is used in many fields (Wallenius et al. 2008). For our problem, it is a consistent choice in the sense it can capture decision maker's perception of preferences. Basically, MAUT defines a model with functions for mapping a utility value for each attribute, and this for each criterion. Then, a weight is assigned to these criteria and a composition function is chosen to aggregate utilities. Once an aggregated utility is computed for each alternative, we can easily establish preferences among them. A performance matrix is then used to answer the respective questions.

Fuzzy sets are interesting when we need to translate these preferences into functions. Introduced by Zadeh (Zadeh 1965), fuzzy sets form the basis of multi-valued membership functions. In crisp traditional logic, the truthfulness

of a statement is a dichotomy, i.e., a statement can only be *true* or *false*. On the other hand, with fuzzy sets, the real unit interval $[0, 1]$ can be used to assign the degree of truth of a statement. The smallest value of the interval denotes a total lack of membership of a class and the highest value denotes a full membership.

To illustrate this concept, let x ($x \in X$) be the financial penalty a suspect is expected to pay ($x \subset \mathbb{N} \,|\, x <= 100\,000, \forall x \in X$). If the suspect is guilty, the judge has to determinate the degree of implication of the suspect (for the sake of brevity, we assume that the penalty depends only on the degree of implication and simplify the procedure). For that, he might only consider a restricted set of linguistic variables: partially implicated, mainly implicated and fully implicated (assuming these alternatives are ordinal). These alternatives can be respectively represented by membership functions: $\mu_{part}(x)$, $\mu_{main}(x)$ and $\mu_{full}(x)$. Membership functions will map these qualitative statements to numbers for each alternative (namely x, the damage to pay), and vice versa. These values are then denoted by fuzzy sets (N being the number of linguistic variables): $\langle x, \mu_1(x), \mu_2(x), ..., \mu_N(x) \rangle$ (when a full partition is considered, $\sum_{i=1}^{N} \mu_i(x) = 1$). The main characteristics of fuzzy sets lie in the choice of the membership functions and the aggregation technique.

Several ideas combine fuzzy logic with MCDM: an interesting summary of this approach is presented by Kahraman (2008). Straccia (2009) presents how to embed fuzzy description logic in MCDM with the perspective to use ontologies. But there are many others: for example, personnel selection in a human resources problem is tackled with a TOPSIS method and an ordered weighted average (in a study undertaken by Dursun and Karsak (2010)); the best strategy for oil spill problems is determined by taking into account several experts' opinions with fuzzy evaluation methods (Krohling and Rigo 2009).

6.2.2 Computational Forensics and Crime Analysis

Computational forensics (CF) is about applying computational methods from several disciplines in the forensic domain. Franke and S. Srihari (2008) defines three axes in which these methods support forensic sciences: (a) they provide a set of tools to overcome limitations of human cognitive abilities; (b) very large sets of data are potentially usable for analyses and are not anymore constrained by the human capacities; and (c) human expert knowledge can be modeled and made explicit to be used in inference mechanisms.

With the support of CF, crime analysis becomes an interesting research field with the goal to overcome traditional limitations. Crime analysis is defined as

> "the systematic study of crime and disorder problems as well as other police-related issues (including socio-demographic, spatial, and temporal factors) to assist the police in criminal apprehension, crime and disorder reduction, crime prevention, and evaluation" (Boba 2009).

Even considering this broad definition, detecting similarities with computational methods between events in a crime analysis perspective has not particularly been an active research area over the past years. Nevertheless, a few researchers have provided some answers. A particular emphasis has been given to analyzing spatiotemporal similarities (such as for predicting serial killers' home (Canter et al. 2000), or for finding a pattern of offenses of serial rapists (J. Warren et al. 1998)). Another focus has been directed towards behavioral crime linkage (e.g., Oatley, Zeleznikow, and Ewart 2005; Santtila et al. 2008; Bennell and Canter 2002; Melnyk et al. 2011; Tonkin, Grant, and Bond 2008; Hazelwood and J. I. Warren 2004) relying mainly on the crime method (the modus operandi) inferred from crime scene investigation.

Even though there exist techniques and methods related to serial crime analysis, some critical questions remain: Can we really detect series in a data set? Do series really exist? A track attempting to provide answers to these questions is based on the Ratcliffe's 4P model (Ratcliffe 2009): *prevention* requires *proactivity*, which requires *predictability* which in turn requires *patterns*. So can we infer that some patterns are present in the data? In the context of serial crimes, our assumption is yes, when we assume that most of the crimes are committed by very few offenders (Wolfgang, Figlio, and Sellin 1987), and moreover when these latter leave a genuine "signature" (Hazelwood and J. I. Warren 2004). In some other cases, situational crime approaches (Cohen and Felson 1979) also suggest a positive answer to our question given that crime occurrences are not randomly distributed and are highly correlated with the physical, spatial, and temporal environment. Repeat victimization phenomena (Weisel 2005) and hot spot locations (Sherman, Gartin, and Buerger 1989) are also factors increasing the probability of patterns.

A recent study (Grossrieder, Albertetti, Stoffel, and Ribaux 2013b) confirms a certain amount of predictability in certain crimes, but a gap still exists. The current study intends to practically support these theoretical assumptions.

Furthermore, a critical overview of the existing crime linkage systems is presented in Bennell, Snook, et al. (2012). A famous but controversial[1] system

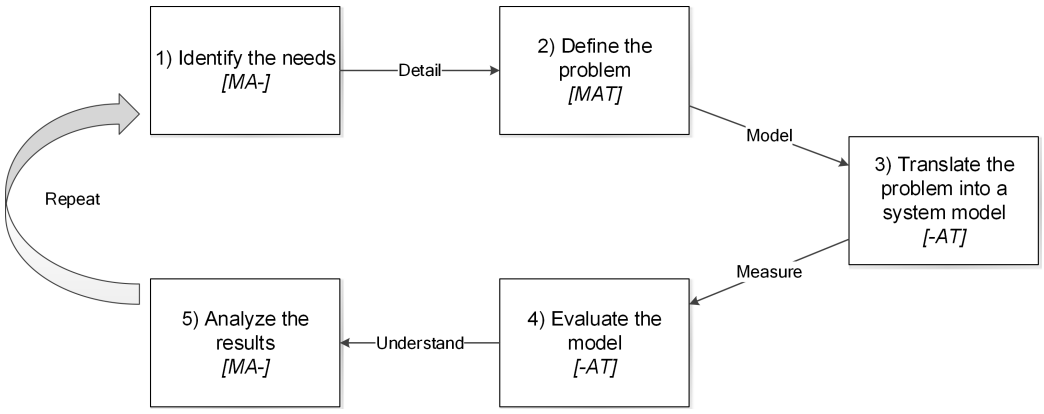


Figure 6.1: The 5 iterative steps of the CriLiM methodology. The *[MAT]* notation indicates for each step the required key competences. "M" stands for *managerial* competences, "A" for *analytic* competences (i.e., crime analysts), and "T" for *technical* competences (e.g., statistical, computational or legal).

(ViCLAS) is assessed, and four assumptions for an effective implementation are formulated based on this analysis. As stated by Bennell et al., "most of the issues discussed in this article apply to all linkage systems". Therefore, these hypotheses are considered in the proposed methodology.

6.3 The Proposed CriLiM Methodology

The proposed Crime Linkage Multi-criteria methodology (denoted hereafter by *CriLiM*) is a set of sequential iterative steps defining a coherent framework for crime linkage. Handling behavioral, situational, and forensic link analysis, the methodology is intended for crime analysts. The idea is to let them express their own perceptions of crime linkage in their own application domain, in order to implement a system tailored to their specific needs.

The problem of similarities is addressed with MAUT methods and fuzzy sets. Using a MAUT method enables the use of different criteria as input to define the notion of similarity, whereas fuzzy sets allow to express the experts' preferences with membership functions.

The conceptual steps of CriLiM are based on the work of Checkland (2000). In a thirty year retrospective of soft systems methodology, he undertakes the

[1] As stated in Snook et al. (2012), despite the fact that ViCLAS exists for nearly 20 years, "there is no published account of its effectiveness being evaluated systematically".

challenge to "see if systems ideas could help us tackle the messy problems of "management" [...]". Actually, a managerial incentive is a requirement in such cases. Many aspects can impact on the level of data quality and on the availability of the data set. Therefore, the overall context is of high importance in order to successfully carry out such projects.

Considering the general MCDM approach, Henig and Buchanan (1996) suggest some tracks to follow in order to determine the key success factors of a "good" decision making process. They argue the fact that good processes should not conceal the preferences of the decision makers themselves but give them a certain degree of liberty, which we will try to achieve with CriLiM.

Based on this analysis, the five following steps constitute the foundation of our methodology (see Figure 6.1):

1. identify the needs and the boundaries of the system;
2. clearly define the problem from a business point of view;
3. translate the business problem into a system model;
4. evaluate the model based on expert results;
5. analyze the results.

The remainder of this section describes these five steps and considers case linkage for non-specific crimes.

6.3.1 Step 1: Identify the Needs and the Boundaries of the System

The first step of the CriLiM methodology is to identify the needs and the boundaries of the system. Basically, the implementation of the methodology should be considered as a project. A business leader is designated to ensure that all the steps are undertaken appropriately. The stakeholders are defined (i.e., business experts, technicians, and managerial representatives) and an approximate scope (including time, economic, and purpose aspects) is given. The required data and legal authorizations are analyzed in accordance with the given departments and the local laws. A clear definition of who will use the system, for what purpose, and to which extent is formulated.

Table 6.1: *(Top)* Performance matrix of m crimes with n properties against c_{ref} with MAUT functions. *(Bottom)* Scores are functions of the utility, and utilities are defined as functions of properties.

	w_1	w_2	...	w_n
	p_1	p_2	...	p_n
c_1	$u^{ref}_{1,1} = \phi^{ref}_1(p_1(c_1))$	$u^{ref}_{1,2} = \phi^{ref}_2(p_2(c_1))$	...	$u^{ref}_{1,n} = \phi^{ref}_n(p_n(c_1))$
c_2	$u^{ref}_{2,1} = \phi^{ref}_1(p_1(c_2))$	$u^{ref}_{2,2} = \phi_{ref,2}(p_2(c_2))$	...	$u^{ref}_{2,n} = \phi^{ref}_n(p_n(c_2))$
...	...	...	...	...
c_m	$u^{ref}_{m,1} = \phi^{ref}_1(p_1(c_m))$	$u^{ref}_{m,2} = \phi^{ref}_2(p_2(c_m))$	...	$u^{ref}_{m,n} = \phi^{ref}_n(p_n(c_m))$

score of c_i against c_{ref}, $i \in [1, m]$
$s_{1,ref} = \eta(u^{ref}_{1,1}, u^{ref}_{1,2}, ..., u^{ref}_{1,n}, w_1, w_2, ..., w_n)$
$s_{2,ref} = \eta(u^{ref}_{2,1}, u^{ref}_{2,2}, ..., u^{ref}_{2,n}, w_1, w_2, ..., w_n)$
$s_{m,ref} = \eta(u^{ref}_{m,1}, u^{ref}_{m,2}, ..., u^{ref}_{m,n}, w_1, w_2, ..., w_n)$

6.3.2 Step 2: Clearly Define the Problem from a Business Point of View

The goal of crime linkage is supposedly to answer a business problem, e.g., "Are some offenders behaving in the same way? Is one single criminal responsible for many burglaries, considering the given data set? Are these events related? How are victims related to offenders?" If an answer to these questions is identified, the crime analyst would then try to explain the possible causes.

Experts and managers together should possess enough knowledge to properly define the purpose and what exactly is expected from the system. They also should state a way to measure the degree of conformance to these statements, i.e., which criteria have to be fulfilled and to which extent.

The identified information that will be used in the linkage system should also be assessed: data quality and data coding are key success factors to successful implementations.

6.3.3 Step 3: Translate the Business Problem into a System Model

Once the problem clearly framed, the definition of a technical system solving the above-mentioned questions can be undertaken. The purpose of this step being to explicit a system detecting similarities within a data set, we will directly strive to define the basis of such system. Actually, the smallest element of this system is an event, which stems from a police report. This list of reported events denotes a set of crimes of similar type for a certain period and in a certain region (e.g., a list of burglaries committed in 2012 in Switzerland).

In order to understand this system, let us define its foundations. The finite list of crimes $C = \{c_1, ..., c_m\}$ is the set of m elements for which we are interested in finding similarities. Each crime $c_i, \forall i \in [1, m]$ is described by n properties, denoted by the set $P = \{p_1, ..., p_n\}$ (which can also be regarded as a set of functions over C: the property p_j of the crime c_i is the value of the function $p_j(c_i)$). Moreover, a degree of importance, denoted as the *weight*, is attached to these properties. Subjective to the decision maker and his experience, weights can be represented by the vector W ($W = \{w_1, ..., w_n\}$).

To create series from the set C, the following steps are applied:

1. set a crime as a reference c_{ref} to which all other crimes c_i will be compared;
2. compute the utility u for each property p of each crime c_i against the reference crime c_{ref};
3. aggregate the utilities of each crime c_i to get the similarity coefficients s against c_{ref};
4. repeat the first three steps m times;
5. create the series based on thresholds according to the similarity coefficients.

The first step consists in choosing a *reference* crime (c_{ref}) and comparing it to all other crimes c_i ($c_{ref}, c_i \in C$) in regard to their properties. The reason of this step is purely iterative. Among all iterations, c_{ref} will vary from c_1 to c_m.

As each crime may have up to n properties, the second step is to compute the utility for each of these properties. It might be computed in many ways. In our situation, we chose to handle the problem by using the multi-attribute

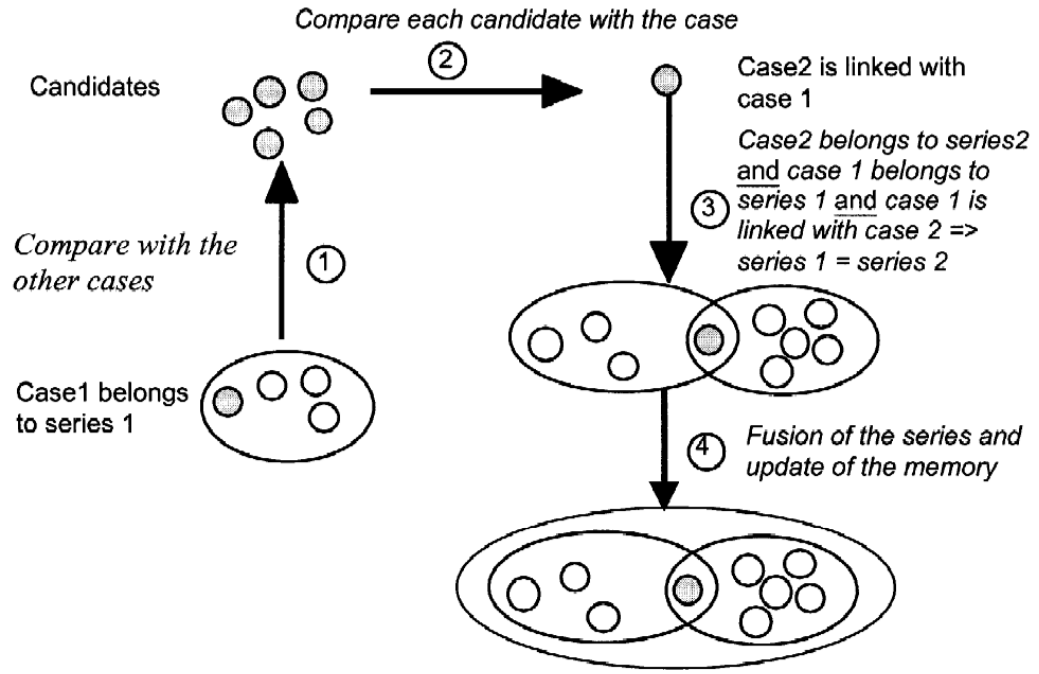


Figure 6.2: Inference structure from Ribaux and Margot (1999) derived from experts: comparing a case.

utility theory technique. MAUT relies on the functions we choose and devise. The idea behind this choice is to give a chance to the experts to express their preferences with the help of fuzzy sets and membership functions. A value function is therefore chosen by the expert according to his experience, which may be subjective. However, in the case of crime analysis, the experts conduct such kind of reasoning on a daily basis and have a wide knowledge of the exogenous factors (many decision criteria often lie beyond the system and beyond the simple concept of weights, because of their intrinsic difficulty to be explicitly defined). This method is therefore an appropriate way to handle uncertainties and to integrate subjective decisions into the system.

The utility of the property p_j of the crime c_i against the reference crime c_{ref} is therefore defined by the utility function ϕ ($\phi : C \rightarrow [0, 1]$) :

$$u_{i,j}^{ref} := \phi_j^{ref}(p_j(c_i)), \tag{6.1}$$

which takes its maximum value in $c_i = c_{ref}$. In the context of using a membership function from fuzzy sets, we suggest a Gaussian curve, a generalized bell-shape, a triangular-shape, or a singleton for ϕ.

The third step is to aggregate these utilities in order to have a single value

Table 6.2: Adjacency matrix, representing the similarity $s_{i,j}$ between each pair of crimes (c_i, c_j). The matrix is symmetric ($s_{i,j} = s_{j,i}$), as similarity functions are symmetric.

	c_1	c_2	...	c_m
c_1	$s_{1,1} = 1$	$s_{1,2}$	...	$s_{1,n}$
c_2	...	$s_{2,2} = 1$	...	$s_{2,n}$
...	...	...	...	...
c_m	...	...	...	$s_{m,m} = 1$

denoting the overall similarity of a crime. A weighted sum can be used (e.g., the SAW method) or any other function can be defined according to the constraints. From a generic point of view, the *aggregation function* η ($\eta : [0,1]^{2n} \rightarrow [0,1]$) defines the similarity of the crime c_i with the crime c_{ref}:

$$s_{i,ref} := \eta(c_i) = \eta(u_{i,1}^{ref}, u_{i,2}^{ref}, ..., u_{i,n}^{ref}, w_1, w_2, ..., w_n). \tag{6.2}$$

This aggregation function has to be valid for a numeric or an ordinal property (see Table 6.1). In the case we are facing nominal non-ordinal properties (such as a multi-valued category or a crisp category), we suggest computing a coefficient regrouping these categories. For this purpose, the Jaccard's index (Jaccard 1908) seems appropriate. The Jaccard's index is a similarity coefficient that has been used in many behavior linkage analysis for modus operandis (Melnyk et al. 2011). For a pair of crimes (c_a, c_b) and crisp memberships of categories, the Jaccard's coefficient of similarity is defined as such:

$$J(c_a, c_b) := \frac{\delta_{a,b}}{\delta_{a,b} + \delta_a + \delta_b}, \tag{6.3}$$

where $\delta_{a,b}$ is the number of categories common to c_a and c_b, δ_a the number of categories that c_a belongs to but not c_b, and δ_b the number of categories that c_b belongs to but not c_a.

The fourth step is to *repeat the same computations, in regard to each crime.* It means we will compute the similarities between all the crimes. From these computations results an adjacency matrix (see Table 6.2), representing the similarity of each pair of crimes. This matrix is symmetric (the similarity is assumed to be reflexive) and the diagonals are equal to 1 (a crime is fully similar to itself).

The last step is to create series in accordance with the adjacency matrix. The method used is based on an inference structure proposed by Ribaux and

Margot (1999), which basically describes crime analysts' implicit reasoning for cases comparison (see Figure 6.2). All pairs of crimes having a similarity above a specific threshold will be merged into a series, depending on the minimum size set for a series. It means that for merging a crime with an existing series, it will be compared to each crimes belonging to the series (all coefficients will have to be above the threshold).

6.3.4 Step 4: Evaluate the Model with Expert Results

The identified series and the adjacency matrix are the output of the model that experts can evaluate. Some specific series should be thoroughly assessed. These series should be clearly presented with all the properties to let the experts visualize the results. Both statistical and visual comparisons —a series can only be evaluated if the offender is actually known— are necessary to evaluate the adequacy of the linkage process in regard to the stated objectives. Evaluation metrics should also be measured. Furthermore, as suggested in Bennell, Snook, et al. (2012), four hypotheses for computerized crime linkage systems should be evaluated: (1) the data provided in the system should be accurate, (2) the data should be coded reliably, (3) patterns should be present in the data, and (4) experts should possess enough knowledge to link crimes accurately.

6.3.5 Step 5: Analyze the Results

The purpose of this step is to analyze the results within the global context, and to identify the subsequent required changes for the next iteration of the methodology.

Basically, the relevance of the results has to be interpreted in regard to the whole project. The use of decision support systems is suggested to properly analyze the results. Pivot tables, dashboards, crime mapping and statistical systems should be considered. In accordance with the results, a list of ad hoc changes is created in order to improve the next iteration.

6.4 Crime Linkage of Residential Burglaries

The goal of this section is twofold: (a) to provide a proof of concept of the proposed methodology, evaluated with the comparison of expert results; and

Table 6.3: Illustration of the burglaries data set (the attribute list is not exhaustive and their values are fictitious). The symbol "-" denotes a missing value.

	Modus operandi 1	Entrance	Phenomenon	Forensic trace	Date	Type of Place	Coordinates
evt248	Grabbed cylinder	Door	Giorno cylindro	-	10-03	Flat	(59,49)
evt994	Tool	Door	Giorno piatto	-	10-04	Villa	(37, 50)
evt093	Climbing	Roof	Notte sera	ID-3095	10-04	Villa	(97, 20)
evt605	Tool	-	-	-	10-04	Flat	(34,32)
evt606	Grabbed cylinder	Door	Giorno cylindro	-	05-02	Flat	(39, 48)
evt038	Broken window	Door	-	ID-3095	10-05	Cellar	(90, 21)

(b) to illustrate the application of this general methodology with a real case study.

6.4.1 Description of the Data

Thanks to the Police de sûreté du Canton de Vaud (a canton police in Switzerland), we were able to access to a data set about serial and itinerant crime (high-volume crime). From about $55,000$ events related to residential burglary involved in 6 cantons for a 4 year period, we focused on a particular subset for this experiment. This subset contains $2,320$ crimes conducted by $1,141$ distinct offenders (during all the study, we assume that each crime has exactly one offender, even if sometimes crimes are committed by several offenders). The distribution of crimes is about 2 per offender on average (with a minimum of 1, a maximum of 52, and a standard deviation of 3.14). A detailed distribution of the series and the relations between the events of the entire data set is presented in Rossy et al. (2013). The reason why we chose this subset is that we only considered the crimes that met some basic requirements to compute a similarity, that is, an offender, a location, a date, and a modus operandi being not null. Two main reasons might explain a missing field: the intrinsic problem of gathering data from a crime scene, or, heterogeneous methods used among distinct police entities to gather data. In order to illustrate the data used for the proof of concept, Table 6.3 represents a fictitious excerpt. Basically, each crime is described with a method of crime, the main entrance used to commit the crime, the type of the place, the GPS coordinates, forensic case data (DNA, shoemarks, earmarks or toolmarks) and a phenomenon type describing the specificity of the crime (a coded value defined by the crime analysts, depending mainly on the modus operandi and the interval of time during the day when the crime occurred). For more information about the nature of the data and how the structure was adapted

from police reports, see Albertetti and Stoffel (2012). The remaining of this section describes step by step how CriLiM is applied to this subset.

6.4.2 Applying the CriLiM Methodology

First, in response to the need for analyzing an increasing amount of data for crime investigation purposes, we focused on a particular problem: the lack of cognitive resources to conduct exhaustive analyses (obviously, crime linkage systems are only a partial answer to this problem). When considering the implementation of some existing solutions, their lack of transparency, their complexity, and their controversy, their implementation may be dissuasive (e.g., the ViCLAS system, as discussed in Snook et al. (2012)). The choice of conducting our own analysis seemed therefore appropriate. Then, the idea was to devise a generic methodology to evaluate the potential of such a system according to our own needs.

Second, we decided to focus on the particular crime of residential burglaries to implement a crime linkage analysis system. An important amount of data was already available and a way to evaluate the results given for most of the crimes, i.e., the offender was known.

The next point was to develop the computerized crime system tailored to the data set (N=$2,320$) and our needs. The result of a first analysis was the need to distinguish utility function types according to the nature of the variables, namely numeric (date and coordinates), categorical (modus operandi, entrance, phenomenon and the type of the place), and unique identifiers (forensic information).

The utility for the first variable type was evaluated with fuzzy sets. For example, the utility of the date property was represented by the fuzzy set $\langle d_i, \mu^{ref}(d_i)\rangle$ (see Figure 6.3). The date was first normalized to the unit interval, denoted by d, and its membership value was evaluated using a bell-shaped function centered on the reference crime date (i.e., $\mu^{ref}(d_{ref}) = 1$).

Then, to compute the score of each pair of events, a simple weighted sum (the SAW technique) was used as an aggregation function for the sake of simplicity. Concerning the weights, a slightly higher importance was given to the date and the location of the crime compared to the other attributes. However, because physical evidence is present in very few cases (11% of N), we decided to add the weighted utility only when the field was not null —meaning that a similarity of 1 can be found even when forensic case data is missing.

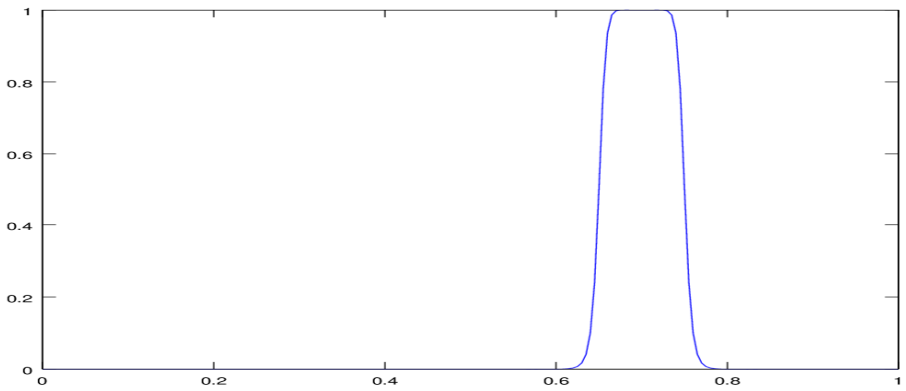


Figure 6.3: Membership function of the date property, represented by a generalized bell-shaped centered on the reference date $d_{ref} = 0.7$ (with parameters of $[0.05, 6, 0.7]$, respectively the width, the slope, and the center of the curve). A full membership denotes a total similarity with the reference date ($\mu^{ref}(d_{ref}) = 1$). An interval of the abscissa of approximately 0.2 denotes a period of 10 months.

The implementation of this model was done using Octave, a language for numeric computations (a free and open-source alternative of MATLAB). The selection of the series is based on the degree of similarity (see the adjacency matrix in Table 6.2) and on the inference structure of comparing cases extracted by experts as described in Ribaux and Margot (1999).

The results of our implementation can be evaluated in many ways. Our choice was to analyze how series were detected by the system (i.e., how crimes were linked together as series when a particular threshold of similarity was reached). We compared these results with 4 different configurations (see Table 6.4), i.e., by changing the minimum coefficient of similarity for linking 2 crimes (the threshold t), and the maximum number of crimes per series (s_{max}). We defined two metrics for each of these configurations: (a) the number of distinct offenders per series, describing the diversity of criminals in a series; and (b), the number of distinct series per offender, counting in how many series crimes of the same offenders are linked.

The most interesting metric is the number of series in which a unique offender is linked. On average, we can see that all crimes of the same offender are split into only one or two series (the best findings occur in the first and the last configuration, i.e., with a number of 10 crimes per series, for a total of respectively 65 and 96 series).

Table 6.4: Results for the implemented solution ($N = 2,320$ crimes). The statistics describe the number of offenders per series (*#o/series*), and the number of series per offender (*#s/offender*). The parameters are the following: t is the similarity threshold, and s_{max} is the maximum size of a series. On average, 890 crimes were linked in at least one series.

	t=0.8, $s_{max} = 10$		t=0.9, $s_{max} = 7$	
	#total linked crimes: 960		#total linked crimes: 894	
Statistics	#o/series ($s_{tot} = 96$)	#s/offender ($o_{tot} = 533$)	#o/series ($s_{tot} = 128$)	#s/offender ($o_{tot} = 452$)
Min	2	1	1	1
Max	10	11	7	11
Mean	7.47	1.35	5.04	1.43
Median	8.00	1.00	5.00	1.00
Std Dev.	2.27	0.94	1.68	1.10

	t=0.95, $s_{max} = 5$		t=0.95, $s_{max} = 10$	
	#total linked crimes: 1060		#total linked crimes: 645	
Statistics	#o/series ($s_{tot} = 212$)	#s/offender ($o_{tot} = 534$)	#o/series ($s_{tot} = 65$)	#s/offender ($o_{tot} = 326$)
Min	1	1	1	1
Max	5	15	10	7
Mean	3.77	1.5	6.63	1.32
Median	4.00	1.00	7.00	1.00
Std Dev.	1.18	1.34	2.40	0.83

These results are encouraging, in the sense that a main objective of a computerized crime linkage system is to give crime analysts a way to reduce linkage blindness. On the other hand, even if crimes of different offenders are linked, it is not a significant issue when considering the concept of false negatives to be more important than the concept of false positives (when the purpose of case linkage focuses on decreasing linkage blindness).

To sum up, the results of such systems are to be used in a context of pure investigation. The main purpose is to enhance cognitive capacities of

analysts, but the conclusions still have to be drawn with extreme care, always considering the assumptions of the model.

6.5 Conclusions

Computerized crime linkage is an emerging research area and many challenges still need to be confronted. In this chapter, we presented a methodology for implementing a tailor-made crime linkage system. The steps of CriLiM were devised to consider the system as a complete solution for experts, instead of providing a single technical tool without context. To cover both high-volume and serious crime, we integrated contextual, behavioral, and forensic information. The approach we adopted derives from fuzzy MCDM in order to deal with similarities articulated with these three dimensions and to express experts' knowledge.

To illustrate and evaluate the methodology, an implementation of CriLiM was applied to residential burglaries. The results illustrated the importance of a good understanding of case linkage systems and its environment, helping in determining the following key success factors: the data quality, the parameters of the system (the thresholds, the size of the series, etc.), and the objectives of the system. More generally, these results have to be drawn with extreme care because they are highly linked to the underlying assumptions of the model.

The impacts of CriLiM for crime analysts may be numerous: provide a basis for implementing their own crime linkage system, provide a method combining several distinct information types for crime linkage, adapt or compare an existing crime linkage system in accordance with the methodology, etc.

With the potential to benefit to all police agencies, CriLiM lays the ground for conducting new experiments and sharing results in the area of crime linkage.

Acknowledgements

The authors would like to thank the Police de sûreté du Canton de Vaud, especially Sylvain Ioset and Damien Dessimoz for their active cooperation and the data provided. We are also grateful to the Swiss National Science Foundation (SNSF) for financial support.

"If you want to understand today, you have to search yesterday."

Pearl S. Buck (1892 – 1973)

7

A Method for Detecting Changes in Crime Trends

The extension of traditional data mining methods to time series has been effectively been applied to a wide range of domains such as finance, econometrics, security, and medicine, to name only a few. Many existing mining methods deal with the task of change points detection, but very few deal with linguistic terms. In this chapter, we propose a method for detecting and querying change points in crime-related time series with the use of a meaningful representation and a fuzzy inference system. Change points detection is based on a shape space representation, and linguistic terms describing geometric properties of the change points are used to express queries, offering the advantage of intuitiveness and flexibility. An empirical evaluation is first conducted on a crime data set to confirm the validity of the proposed method and then on a financial data set to test its general applicability. A comparison to a similar change-point detection algorithm and a sensitivity analysis are also conducted. Results show that the method is able to accurately detect change points at very low computational costs. More broadly, the flexible detection of specific change points within time series of virtually any domain is made more intuitive and more understandable, even for experts not related to data mining.

This essay can be found as "Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation," Albertetti, Grossrieder, et al. (2016).

7.1 Introduction

The analysis of time series naturally arises in crime analysis as well as in any data-driven domain. Finding sudden changes in criminal activities is a particular task known as change points detection. In this paper, a flexible on-line change points detection method for helping crime analysts to easily and understandably monitor changes is proposed. Change points are detected in two steps: the segmentation of the time series and the querying of points with a fuzzy inference system.

7.1.1 Motivation

Knowledge extraction of time series can be viewed as an extension of traditional mining methods with an emphasis on the temporal aspect. Among these, *Change points detection* methods focus on finding time points at which data *suddenly* change (in contrast to *slow* changes). Many studies have shown interesting applications of change points detection in various domains. These methods are based on neural networks, regressions, or other statistical models, with an emphasis on the efficiency of these methods. However, only a few consider approaches with these two properties: a meaningful and expressive subspace representation of the time series, and a dynamic segmentation process without fixed-sized windows, linked together flexibly.

In the domain of crime analysis, such flexible and intuitive approaches for change points detection are particularly sought, especially for crime trends monitoring. Previous studies from the authors ((Albertetti and Stoffel 2012), (Albertetti et al. 2013a), (Albertetti et al. 2013b), and (Grossrieder, Albertetti, Stoffel, and Ribaux 2013b)) emphasize on the usefulness of crime trends monitoring activities and advocate the use of appropriate methods for considering the specificities and the constraints of the crime analysis domain, that is basically dealing with uncertainties. The automated process of change points detection is considered as a major step in the production of intelligence, supporting the activity of crime analysis (also sometimes referred to crime intelligence).

Flexible change points detection methods are critical for supporting analysts in their daily tasks, especially for the monitoring of serial and high-volume crimes (e.g., burglaries). Most of the time, crime analysts have no particular background in time series analysis, but still need to analyze and monitor crime trends. These trends are drawn into the whole activities and are not

always perceived by police forces. As for example, querying criminal activities about a particular increase in crime trends for targeted police interventions, as well as querying patterns of changes for the general understanding of crime phenomena are common tasks.

Finding changes in crime trends assumes two conditions: (a) the actual existence of a trend, and (b) its detection within the data. The first condition is far from obvious, but as crime analysis is founded on environmental criminology theories, a justification for the existence of crime trends appears ((Grossrieder, Albertetti, Stoffel, and Ribaux 2013b), (Boba 2009), and (Felson and R. V. Clarke 1998)). The second is generally simply assumed, but difficult to detect in massive data sets and needs intuitive and understandable analytical methods.

Although the proposed method is a specific answer for the domain of crime analysis, we believe that it has great potential applications in several domains. As an example, in the financial domain, it proves very useful to find and query change points in real-time, giving the investors flexible means to detect trends indicating the right moment for selling or buying stocks.

7.1.2 Contribution of this paper

The method proposed in this paper, for the Fuzzy Change Points Detection in crime-related time series (FCPD), aims to focus on flexibility and intuitiveness. To achieve this purpose, the method combines a segmentation step and a querying step. Moreover, the following characteristics make FCPD unique:

- a meaningful and expressive representation of the time series is used ;
- the segmentation is *dynamic*, that is, segments are set according to the underlying shapes of the time series, without using a fixed-size window parameter;
- changes are queried with linguistic terms, using a fuzzy inference system;
- the method does not rely on training sets;
- the method is on-line and iterative, i.e., change points can be detected with past values only and there is no need to compute the entire model at each new observed value. The computational cost is very low; it is related to the size of the approximating polynomials (instead of the number of the observations).

Indeed, with the use of a meaningful representation and a dynamic segmentation, change points can be more easily described and identified. The segments found in a time series, reflecting change points, are described with meaningful estimators such as the average, the slope, the curvature, etc. Then, with the use of a fuzzy inference system, a query can be specified using linguistic terms describing the geometric estimators. It becomes then easier, for instance, to query a time series about the most abrupt changes in terms of slope. In the example "IF *average* is *low* AND *slope* is *very_high*, THEN *pertinence* is HIGH", the inference system would return a high score on segments representing shapes of the given description. This approach makes the querying of change points particularly intuitive and flexible, especially for domain experts.

7.1.3 Structure of this paper

The remainder of this paper is structured as follows: in Section 7.2, a literature review in the mining of time series is provided; Sect. 7.3 introduces some concepts in the preparation, representation, and analysis of time series; Sect. 7.4 details FCPD, a step-by-step method for the fuzzy querying and detection of change points in crime-related time series; an empirical validation on synthetic and real-world data is conducted in Sect. 7.5; results are discussed in Sect.7.6; and finally in Sect. 7.7 a conclusion is drawn from the experiments and some tracks for future work are suggested.

7.2 Literature review

Change points detection has numerous application domains, as for example finance, biology, ocean engineering, medicine, and crime analysis. It is considered as a final objective in the whole process of time series analysis among classification, rules discovery, prediction, and summarization. Almost all of these mining tasks require data preparation, namely the representation of the time series, its indexing, its segmentation, and/or its visualization. In this section, we propose a review of these steps, before comparing existing methods for change points detection. An extensive review of the analysis of time series can be found in Fu (2011), as well as a general methodology in Last, Klein, and Kandel (2001).

7.2.1 Representation of time series

Many representation models of time series have been dealt with in the literature, each claimed with relative advantages and drawbacks. Two main categories are symbolic representations and numeric representations. Symbolic representations are less sensitive to noise and are usually computationally faster. For the last decade, the community has been paying particular attention to the Symbolic Aggregate Approximation (SAX) representation ((Lin, E. Keogh, Lonardi, et al. 2003) and (Lin, E. Keogh, Wei, et al. 2007)), with the main advantages to reduce the original dimensionality of the data, being on-line, and having a robust distance measure. However, it does not cover all needs. In Fuchs, Gruber, Pree, et al. (2010), a numeric representation —which differs from SAX and many others by giving a meaning to the representation— is used to perform several mining tasks. This *shape space* representation uses coefficients as shape estimators of the time series it represents, leading to an intuitive description.

7.2.2 Segmentation of time series

Most mining methods use subsequences (or segments) of time series as input to the analysis. Segmentation algorithms with the approach of a sliding window are simple to use but present the main drawback of being static, i.e., segmenting the time series according to a fixed and exogenous parameter (e.g., the length of the window) without considering the observed values. Other algorithms, based on a bottom-up or a top-down approach are considered as dynamic (e.g., by using some error criteria as segmentation thresholds) but need the whole data set to operate. These off-line algorithms usually perform better in terms of accuracy but have higher computational costs and are not suitable for real-time applications. A combination of the aforementioned algorithms, namely the SWAB segmentation algorithm, is presented in E. J. Keogh et al. (2001). A study (E. Keogh et al. 2004) provides benchmarks on these claims and as a result suggests that SWAB is empirically superior to all other algorithms discussed in the literature. As we believe there is no silver bullet, each application has its own requirements. A more flexible approach is the SwiftSeg algorithm (Fuchs, Gruber, Nitschke, et al. 2010), providing a dynamic and on-line approach to segmentation, with the possibility of a mix between growing and sliding window. Another interesting segmentation approach (F.-l. Chung et al. 2002), specific to stock mining and described as dynamic, is based on the identification of perceptually important points (PIP).

7.2.3 Fuzzy analysis of time series

A small subset of temporal mining methods takes advantage of the characteristics offered by fuzzy logic and fuzzy sets. The concept of fuzzy time series has first been defined by Song and Chissom in Song and Chissom (1993a) and Song and Chissom (1993b), with an application in class enrollment forecasting. Soon followed multiple variations and improvements of the basic method (e.g., (S.-M. Chen 1996), (Hwang, S.-M. Chen, and C.-H. Lee 1998), (Huarng 2001), or (C.-H. Chen, Hong, and Tseng 2012)), with their own types of fuzzy inference systems (FIS). Two common FIS, namely the Mamdani inference system (Mamdani 1974) and the Takagi-Sugeno inference system (Sugeno and Takagi 1985), can be intuitively used to deal with uncertain and flexible data. In C.-H. L. Lee, A. Liu, and W.-S. Chen (2006), an application in finance uses an FIS for pattern discovery. In Güner and Yumuk (2014), prediction of long shore sediments is also dealt with the use of an FIS. In parallel, a combination of FISs and neural networks have found an origin in Jang (1993). As for examples, the prediction of time series is performed with dynamic evolving neuro-fuzzy inference systems (Song and Kasabov 2000), the classification of electroencephalograms (Güler and Übeyli 2005), as well as the prediction of hydrological time series (Zounemat-Kermani and Teshnehlab 2008).

7.2.4 Change points detection

Change points detection in time series analysis has been thoroughly investigated, mainly using statistical models (see Basseville and Nikiforov (1993) for a general introduction). Reeves et al. (Reeves et al. 2007) attempt to review and compare the major change points detection methods for climate data series.

More specifically, related approaches for change points detection have been investigated in a relatively limited set of studies. For example, a statistical based approach using fuzzy clustering is described in B. Wu and M.-H. Chen (1999) and Kumar and B. Wu (2001). Verbesselt et al., in Verbesselt, Hyndman, Newnham, et al. (2010) and Verbesselt, Hyndman, Zeileis, et al. (2010), detect breaks for additive seasonal and trends (BFAST), with a principal application is phenology. To deal with imprecise observation in time series, changes are analyzed with fuzzy variables in Cappelli, D'Urso, and Iorio (2013). In X. C. Chen et al. (2013), a contextual change detection algorithm addresses relative changes with respect to a group of time series. In Yamanishi and Takeuchi (2002) and Takeuchi and Yamanishi (2006), the utility of a

framework for outliers detection of time series prediction is highlighted. In C.-H. Chen, Hong, and Tseng (2012), the need to use linguistic values for comprehensible results is advocated, where fuzzy time series mining is used for association rules between data points (but not between segments) with fixed-size window. A qualitative description of multivariate time series with the use of fuzzy logic is presented in Moreno-Garcia et al. (2014). Yu et al. (Yu, Tzeng, and H.-L. Li 2001) propose a fuzzy piecewise regression analysis with automatic change points detection. In H. Wang, D. Zhang, and Shin (2004), DoS attacks are monitored with a change point approach based on the non-parametric Cumulative Sum (CUSUM) method.

7.3 Time series representation and fuzzy concepts

In the following subsections, a review of a time series representation using polynomials is presented and their main advantages are explained. Then a dynamic segmentation method is described. Finally, concepts of fuzzy time series and fuzzy inference systems, which are useful to analyze segments, are introduced.

7.3.1 A polynomial shape space representation

Let us consider a time series defined by the sequence

$$s = (y_0, y_1, \ldots, y_N), \quad y_i \in \mathbb{R} \quad (i = 0, 1, \ldots, N) \tag{7.1}$$

of $N + 1$ points measured over the equidistant points in time $x_0, x_1, \ldots, x_N$. Basically, our set of points s can be modeled by a parametrized function $f(x)$, which is obtained with a linear combination of basis functions f_k:

$$f(x) = \sum_{k=0}^{K} w_k \, f_k(x) \, . \tag{7.2}$$

The properties of this approximation depend on the choice of the basis functions f_k and their weights. Given some appropriate basis functions, an optimal approximation can be found with the vector of weights $\mathbf{w}^* \in$

$\mathbb{R}^{K+1}, \mathbf{w}^* = (w_0, w_1, \ldots, w_k)^T$, which minimizes the approximation error in the least-squared sense. Fuchs et al., in Fuchs, Gruber, Nitschke, et al. (2010), claim that these weights show interesting properties when using some specific of these $K+1$ basis functions. Indeed, when particular conditions are met, these weights describe the shape of the considered time series intuitively. As a corollary, an efficient similarity measure can be defined based on the extracted features.

Let us now describe these particular approximating polynomials, as in Fuchs, Gruber, Nitschke, et al. (2010), with

$$p(x) = \sum_{k=0}^{K} \alpha_k \, p_k(x) \,, \tag{7.3}$$

where $p(x)$ is the approximating polynomial, the polynomials p_k are the basis functions f_k and the coefficients α_k are the weights w_k, relating to Equation 7.2. These coefficients are defined as

$$\alpha_k = \frac{1}{\|p_k\|^2} \sum_{n=0}^{N} y_n \, p_k(n) \,, \tag{7.4}$$

where

$$\begin{aligned} p_{-1}(x) &= 0 \,, \\ p_0(x) &= 1 \,, \\ p_{k+1}(x) &= (x - a_k) \, p_k(x) - b_k \, p_{k-1}(x) \,. \end{aligned}$$

Then, by defining $\boldsymbol{\alpha}$ as the *vector of coefficients*

$$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_K)^T \,, \tag{7.5}$$

any time series can be characterized by these coefficients only, given some polynomials. The interesting property is that the i_{th} coefficient represents the i_{th} derivative of the approximated time series; i.e., the first coefficient α_0 is an optimal estimator (in the least-squared sense) for the average of the considered $N+1$ data points, α_1 an estimator for the slope, α_2 an estimator for the curvature, etc.

The parameter K of the orthogonal expansion (Eq. 7.3) has to be carefully chosen in accordance with the desired description of the time series. As depicted in Figure 7.1, setting $K = 0$ defines a single polynomial term in Eq. 7.3 with a maximum degree of 0 and a corresponding vector coefficient $\boldsymbol{\alpha} \in \mathbb{R}^1$, representing a time series according to its average only; setting $K = 1$ adds the estimator of slope; setting $K = 2$ adds on top the estimator of curvature; and so on. Choosing this parameter is a trade-off between computational costs and representation accuracy.

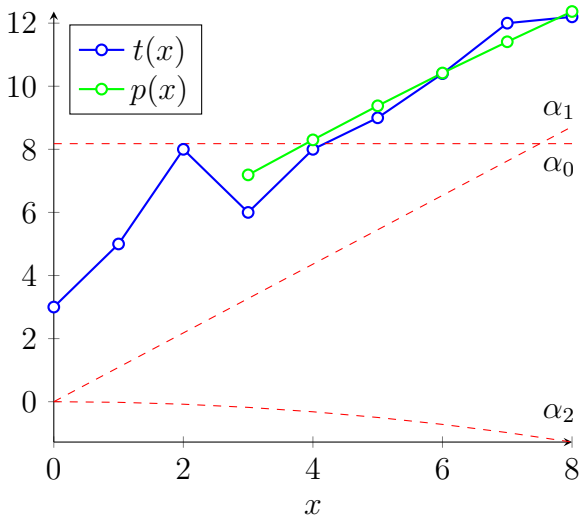


Figure 7.1: A time series $t(x)$ and its polynomial approximation $p(x)$ with $K = 2$, and their coefficients. The coefficients $\alpha_0 = 8.18$ (estimator for the average), $\alpha_1 = 1.09$ (estimator for the slope), and $\alpha_2 = -0.02$ (estimator for the curvature) are depicted with the first term only of their respective polynomials. It should be noticed that the approximation can only start from $K + 1$ data points.

In order to hold these desired properties, each of the given polynomials p_k defining the *orthogonal expansion* of the approximation must:

- have different ascending degrees 0,1,...,K;
- have a leading coefficient of 1;
- be orthogonal with respect to the inner product

$$\langle p_i | p_j \rangle = \sum_{n=0}^{N} p_i(x_n)\, p_j(x_n), \quad i \neq j\,. \tag{7.6}$$

For instance, the discrete Chebyshev's polynomials reveal these criteria. Defined in a recursive way, the first Chebyshev's terms are:

$$\begin{aligned}
p_0(x) =& 1, \\
p_1(x) =& x - \frac{N}{2}\,, \\
p_2(x) =& x^2 - Nx + \frac{N^2 - N}{6}\,, \\
p_3(x) =& x^3 - \frac{3N}{2}x^2 + \frac{6N^2 - 3N + 2}{10}x \\
& - \frac{N(N-1)(N-2)}{20}\,.
\end{aligned}$$

More generally, a term of the series is defined as

$$p_{k+1}(x) = \left(x - \frac{N}{2}\right)\, p_k(x) - \frac{k^2((N+1)^2 - k^2)}{4(4k^2 - 1)}\, p_{k-1}(x)\,, \qquad (7.7)$$

and their squared norms are given by

$$\|p_k\|^2 = \frac{(k!)^4}{(2k)!\,(2k+1)!} \prod_{i=-k}^{k} (N + 1 + i)\,, \qquad k = 0, 1, \ldots, K\,. \quad (7.8)$$

Based on these definitions, it is now possible to redefine our time series s from Eq. 7.1. As the vector $\boldsymbol{\alpha}$ contains the estimators up to degree K for the considered time series, it is said that s *is approximated by* $\boldsymbol{\alpha}$ with the statement

$$s = (y_0, y_1, \ldots, y_N) \sim \boldsymbol{\alpha}\,, \qquad (7.9)$$

where $s \in \mathbb{R}^{N+1}$, $\quad \boldsymbol{\alpha} \in \mathbb{R}^{K+1}$, $\quad K \ll N$.

7.3.2 Segmenting time series

Segmenting time series is useful for analyzing and comparing subsets of data points. The considered segmentation approach in this paper is *dynamic*, meaning segmentations are performed in accordance with the intrinsic shapes underlying in the data, in contrast to other segmentation approaches that only depend on an artificial window size or with equal size segments. To do so, the sequence s from Eq. 7.1 is split into the set of contiguous windows

$$W(s) = \bigcup_{s^{(i)} \in s} s^{(i)}, \quad i \in \{0, 1, \ldots, N/2\}, \tag{7.10}$$

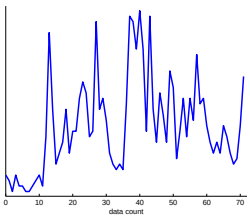
where W is a partition of s, and $s^{(i)}$ is the i_{th} segment containing two or more elements. To be dynamic, the partitioning is done by detecting *change points* $\hat{\boldsymbol{x}}$ within our time domain $x = (x_0, x_1, \ldots, x_N)$, relating to the concept of abrupt changes detection (Basseville and Nikiforov 1993).

Let us consider a small example. Within the sequence $s = (y_0, y_1, \ldots, y_{10})$, the change points $\hat{x}$ detected are x_3 and x_7; then $\hat{\boldsymbol{x}} = (3, 7)$, $W(s) = \{s^{(1)}, s^{(2)}, s^{(3)}\}$, with $s^{(1)} = (y_0, y_1, y_2, y_3)$, $s^{(2)} = (y_4, y_5, y_6, y_7)$, $s^{(3)} = (y_8, y_9, y_{10})$. It has to be emphasized that these contiguous windows do not need to have the same size. In fact, their underlying estimators $\boldsymbol{\alpha}$ only depend on the shape of the segment. Therefore, *primitives*, or *basic* shapes are more accurately represented by these estimators, as the deviation of the predicted values (the sum of the residuals) is low within the considered segment (i.e., the estimators do not significantly change within a segment while the windows is growing). This property justifies why an adaptive way of segmenting the time series with change points is preferred to fixed-size window segmentation methods, considering the objective of this study.

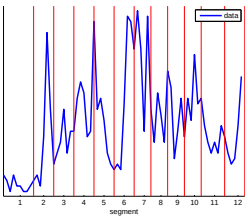
As part of an iterative process, the first segmentation step starts with a window from the first point of the segment, setting $x = 0$. The corresponding orthogonal expansion is computed, holding the $\boldsymbol{\alpha}$ coefficients. Then, specific criteria based on the coefficients are derived (such as the deviation of the predicted value or the count of sign switches of the slope) and compared to thresholds. If the thresholds are exceeded, the growing process is stopped, a change point is detected and a new segment starts with the next available data point; otherwise, the window keeps growing to the next point, the expansion is updated and the new coefficients are again compared. Figure 7.2 depicts an example of this segmentation.

This segmentation method is based on the *SwiftSeg* algorithm from Fuchs et

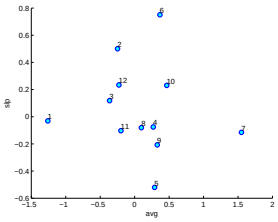
Figure 7.2: Illustration of the segmentation of a time series with its shape-space representation.



(a) A real time series of 72 samples. The time series has been normalized ($\mu = 0$, $\sigma^2 = 1$). From the pretty chaotic shape of the series, many change points are supposed to be found.



(b) Segmentation of the time series, with $K = 3$. Twelve different segments are found (represented by vertical lines). The segmentation here is based on the number of sign changes of the slope and the deviation of the predicted value.



(c) Segments depicted with their shape space representation (only the first two coefficients are shown). The segments 1 and 7 are easily identified as with the lowest and highest average (α_0) and the segments 5 and 6 with the lowest and highest slope (α_1).

al. (Fuchs, Gruber, Nitschke, et al. 2010). Their study describes an on-line algorithm for updating the values of the coefficients, where the computation only depends on the last point added to the window, leading to effective computational costs; in contrast to off-line algorithms that need the entire window to update the coefficients. Combinations of growing and fixed-length window are also documented and experimented.

7.3.3 Fuzzy time series

The concept of fuzzy time series as first defined by Song and Chissom in Song and Chissom (1993a) is here resumed. Let us consider the universe of discourse

$$U = \{u_1, u_2, \ldots, u_m\} \tag{7.11}$$

and the set

$$A_i = \mu_{A_i}(u_1)/u_1 + \ldots + \mu_{A_i}(u_m)/u_m\,. \tag{7.12}$$

A_i is a fuzzy set of U, where '/' indicates the separation between the membership grades and the elements of the universe of discourse U, '+' is the union of two elements, and the fuzzy membership function

$$\mu_{A_i}(u_j) : U \rightarrow [0, 1] \tag{7.13}$$

expresses the grade of membership of u_j in A_i.

Let the elements of our time series $(y_t)(t = 0, 1, \ldots, N)$, a subset of $\mathbb{R}$, be the universe of discourse replacing U on which the fuzzy sets $A_i(i = 1, 2, \ldots)$ are formed and let f_t be a collection of $\mu_{A_i}(t)(i = 1, \ldots, m)$. Then, $f_t(t = 0, 1, \ldots, N)$ is called a fuzzy time series on $y_t(t = 0, 1, \ldots, N)$.

A fuzzy relationship between one point at time (t) and its successor is represented by:

$$f_t \implies f_{t+1}\,. \tag{7.14}$$

We suggest a slightly more generic definition that can deal with segments. Indeed, we will consider fuzzy relationships between *any element* at time (t) and its successor, with f_t being the segment $s^{(t)}$ and f_{t+1} its segment $s^{(t+1)}$ (i.e., f_t describes the *entire* segment, instead of a specific point of the time series).

7.3.4 Fuzzy inference systems

Fuzzy inference systems (FIS) can model uncertain and complex human reasoning tasks. FISs use "IF antecedent THEN consequent" rules as inference mechanism, where the antecedent and the consequent of the rule are linguistic terms that can handle multi-valued logic.

Different types of fuzzy inference systems exist. Two of them are widespread in the literature, namely the Takagi-Sugeno (Sugeno and Takagi 1985) and the Mamdani (Mamdani 1974) type. The main difference between these two is that the latter uses output membership functions to describe linguistic terms, whereas the former uses output membership functions to describe crisp values. In this paper, the Mamdani inference system is considered because of its relative simplicity.

A fuzzy inference system is defined (see Fig. 7.3) by a *rule base* containing the set of "IF-THEN" rules; a *database* with the fuzzy sets and their membership

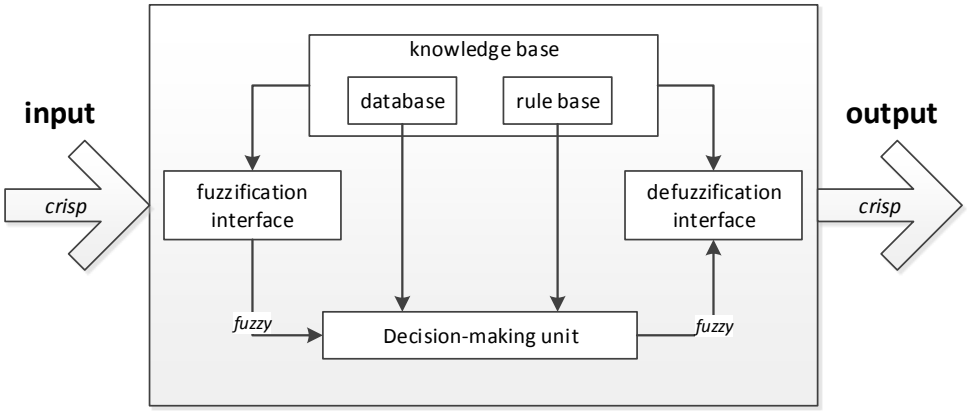


Figure 7.3: Structure of a fuzzy inference system.

functions; a *decision-making unit* performing inference based on the rules; a *fuzzification interface* transforming the crisp inputs into degrees of match with linguistic values; and a *defuzzification interface* transforming the fuzzy results of the inference into numbers.

On top of this structure, the inference process is defined according to the 5 following stages:

1) fuzzification of the inputs;

2) combination of the antecedents with conjunction or disjunction functions;

3) rules firing and implication of the consequent;

4) aggregation of the consequent; and

5) defuzzification of the output.

7.4 The proposed method for the fuzzy detection of change points in crime-related time series

In this section, a novel method for the Fuzzy Change Points Detection in crime-related time series, FCPD, is proposed. Based on the particular shape

space representation and the dynamic segmentation method described in Sect. 7.3, a unique approach is proposed.

FCPD consists of 2 steps:

1) the *segmentation* of the time series by the means of the shape space representation (represented in pseudo-code by Algorithm 1);

2) the fuzzy *querying*, by the use of linguistic variables, of change points based on the discovered segments (represented in pseudo-code by Algorithm 2).

These 2 steps are performed on-line. Algorithm 1 starts with the very first observation of a time series and grows a window at every new observation. Every time a new segment is set (defined by some error criteria), Algorithm 2 can be run to answer queries bases on the discovered segments (i.e., the outputs of Alg. 1). The FIS structure (i.e., the membership functions, the linguistic variables and the rules) and the query are defined by the user in accordance with the use case and do not change over time.

Algorithm 1: SEGMENTATION, for on-line segmenting

```
input  : s = (y_0, y_1, ..., y_n), the time series
         K, the degree
         th, the thresholds
output : α, the coefficients
begin
    w ⟵ initWindow(s[0..K])
    c ⟵ initCoefs(w, K)
    i ⟵ K + 1
    while i <= n do
        w ⟵ growWindow(w, s[i])
        c ⟵ updateCoefs(c, w, K)
        if newSegment(c, th) then
            remove s[0..i] from s
            add c to α
            β ⟵ query(α)
            goto begin
        i ⟵ i + 1
```

Algorithm 2: QUERY, for on-line change points querying

input : q, the query (global variable)
F, the FIS structure (global variable)
$\boldsymbol{\alpha}$, the coefficients
output : $\boldsymbol{\beta}$, the sorted segments
begin
$FIS \longleftarrow initFIS(F)$
$scores \longleftarrow inferFIS(FIS, \boldsymbol{\alpha}, q)$ 'see Fig. 7.3
$\boldsymbol{\beta} \longleftarrow sortSegments(scores)$

Step 1: Finding the segments $W(s)$ of the time series

Starting with a time series represented by $s = (y_0, y_1, \ldots, y_N)$ as in Eq. 7.1, the parameter K (i.e., the degree of the polynomials), and the thresholds, the segmentation process (Sect. 7.3.2) is iteratively applied. First, a growing window is positioned on the first element (y_0), the polynomial expansion is computed and its coefficients $\boldsymbol{\alpha}$ are extracted according to the chosen degree K. Based on these coefficients some segmentation criteria can be defined. Two of them are hereafter suggested.

The first threshold is the deviation of the predicted value:

$$c_{DPU} = \begin{cases} 1, & \text{if } |p(x_t) - y_t| > th_{DPU}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.15}$$

where $p(x_t)$ is the predicted value of the regression, $y(x_t)$ the actual value at time t (the last point of the growing window), and th_{DPU} the value of the threshold. The second is the counting of the sign switches of the slope:

$$c_{SSS} = \begin{cases} 1, & \text{if } SSS > th_{SSS}, \\ 0, & \text{otherwise,} \end{cases} \tag{7.16}$$

where SSS counts the number of sign switches of the slope within the window, that is, SSS is incremented if a change in the sign of the slope is observed, and th_{SSS} the value of the threshold. A new segment $s^{(i)}$ is added to $W(s)$ if these criteria are met (one single criterion can be enough), and the segment is then represented by the last $\boldsymbol{\alpha}$ computed, $\alpha^{(t)}$. Otherwise, the window is grown by adding the next point and the same steps are repeated until the end of the time series (i.e., $\boldsymbol{\alpha}$ is updated and the new criteria are again compared to the

same thresholds). The result of this step is the set of continuous windows $W(s)$, their respective coefficients $\boldsymbol{\alpha}^{(i)}$ and their change points $\hat{x}^{(i)}$.

Step 2: Querying the existence of particular change points with a fuzzy inference system

A query consists of the expression of geometric properties with linguistic values related to the coefficients, and the result of the query is the corresponding "relevance score" of each segment regarding the query. The query is specified with a fuzzy "IF-THEN" rule. To evaluate the query, features related to the coefficients from the subspace representation are given as input to the FIS, the query is added to the rule base, and the system infers the output. The defuzzification of the output is the answer to the original query. The membership functions of the FIS have to be specified at the beginning. In fact, linguistic variables and membership functions are part of the query, specified by the user to detect change points with regard to their applications. These parameters should not change over time, unless if the query itself changes. Setting the FIS amounts to:

a) *Choosing the input(s) and the output(s) of the fuzzy inference system in relevance of the query.*
Inputs constitute the antecedent part of the rules and outputs the consequent. As initial intuition, inputs could be the coefficients $\boldsymbol{\alpha}^{(i)}$. The use of these coefficients enables the expression of linguistic terms concerning the average, the slope, or the curvature of the segments in the antecedent part of the rules.

However, to handle queries considering more aspects, other input variables can be considered to get different expressions. Relations between two *elements* (as in Eq. 7.14, where $f_t \implies f_{t+1}$), can also be inputs. For instance, if the element is a segment, the input of the FIS is then the coefficient variation between two segments, defined as:

$$v_{\alpha_k}(s^{(t)}, s^{(t+d)}) = \frac{\alpha_k^{(t+d)} - \alpha_k^{(t)}}{\alpha_k^{(t)}}\,, \tag{7.17}$$

where $s^{(i)}$ is the segment of index i in $W(s)$, $\alpha_k^{(i)}$ the coefficient of order k of this segment, and d the delay operator of segments, typically set to 1. For the sake of simplicity, these variations will be referred to as

$$v_{\alpha_k}^{i->i+d}. \tag{7.18}$$

In other words, the use of variations, instead of the coefficients input to the FIS, enable to express relative changes between two periods instead of absolute changes of value only.

Other inputs to consider are for instance the size of each segment, the variation of the size, or a set of primitive shapes. A combination of these is also possible.

The output of the FIS is more straightforward. Given the rules, the FIS outputs the degree of similarity of each input to the geometric properties specified in the query. Therefore, only one output —the relevance of the query— is assumed to be necessary in most cases. A FIS with several outputs is nonetheless possible.

b) *Defining the linguistic terms and their membership functions.*
Each input/output of the inference system is generally defined by multiple fuzzy sets. For example, (*LOW*), (*MEDIUM*), and (*HIGH*) can be fuzzy sets for the coefficients as input, whereas (*DECREASE*), (*CONSTANT*), and (*INCREASE*) are sets of variations between elements. For these terms we need membership functions that can be valued, as part of the fuzzification and defuzzification process.

c) *Defining the inference rules.*
The rules added to the inference system are the heuristics that guide the search to find the appropriate change points. These heuristics use the geometric estimators from the coefficient to express visual criteria of the researched segments. These inputs are evaluated in the antecedent of the rule and the output in the consequent. A weight can be added to each rule, giving different degrees of importance in accordance with the confidence of the heuristic.

d) *Inferring the output(s).*
Infer the output(s) of the FIS (as described in Sect. 7.3.4). According to the rules, the segments which are the most relevant to the query output a higher membership of the consequent.

7.5 Empirical Evaluation

This section provides an empirical evaluation of the proposed method, with a focus on crime data. For the sake of an overall evaluation, different types of time series with different objectives are analyzed.

First, qualitative analyses, representing illustrated case studies helping practitioners to better evaluate the method, are conducted (with a total of 4 time series):

1) the analysis of cyclic data, to illustrate the use of the proposed method in a simple environment;

2) a case study of crime trends monitoring, to support the validity and applicability of flexible change points detection and querying according to the domain of crime analysis;

3) the analysis of the *TOPIX* time series (financial real-world data), in a financial case study, to test the domain-free applicability of FCPD;

Second, quantitative analyses, each time systematically compared with two comprehensive data sets (with a total of 96 time series):

4) a comparison with a similar change points detection algorithm, BFAST, is performed on both the CICOP and the SWX data sets for assessing the accuracy and the complexity of FCPD;

5) a sensitivity analysis on both the CICOP and SWX data sets is carried on for measuring the impact of the parameters on the results of the proposed method.

These two data sets used in the quantitative part are the following:

A) the *CICOP* data set (crime real-world data), consisting of 32 time series each with 70 observations. These time series describe monthly events for a period of 6 years (2009-2014) of serial- and itinerant-related crimes (such as burglaries);

B) the *SWX* data set (financial real-world data), consisting of 64 time series each with 120 observations. These time series describe monthly stock data for a period of 10 years (2005-2014) from the small and medium capitalizations of the SWX (Swiss Exchange Market).

For these experiments, a MATLAB version of FCPD has been implemented by the authors. Time series were normalized ($\mu = 0$, $\sigma^2 = 1$) before analysis. The only reason for normalizing time series is to make comparison between

different data sets easier: indeed, normalizing time observations leads to consistent thresholds throughout all time series (i.e., thresholds in accordance with the mean and the variance of the time series).

7.5.1 Simulation with cyclic data

This first experiment aims to detect simple change points within the cycle time series. For that purpose, we generated a normalized sinusoidal time series of 2000 data points, representing a cyclical activity. We introduced two "anomalies", the first between the x-interval $[500, 600]$ by adding noise with a standard normal distribution ($N(0, 1)$) to the observed values and the second between $[1400, 1600]$ by replacing the observed values with $y = 0.5 + u/2$, where u is a noise factor with a standard normal distribution.

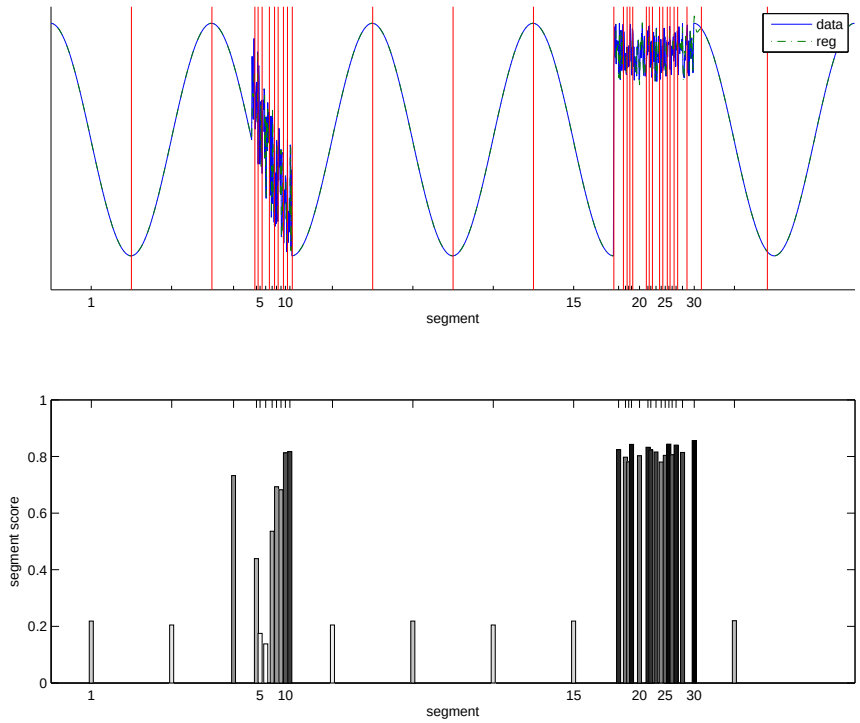


Figure 7.4: The cycle time series. *(Top)* Segmentation of the time series, 31 change points detected. *(Bottom)* Output of the FIS for each segment, representing the relevance of the query.

For the segmentation step, we set the maximum degree of the polynomial regression K equals to 5 and a single threshold for the switches of the sign slope (th_{SSS}) to the value of 1. As a result, 31 change points were detected (top of Fig. 7.4). We then used as input to the FIS the coefficient matrix for the average (i.e., the coefficients of degree 0 for each segment). The input is described with 3 different linguistic terms, namely *negative*, *zero*, and *positive*, respectively represented by a Z-shaped, a Gaussian, and a S-shaped membership function as depicted in Fig. 7.5. The output membership is a triangular-shaped function, using *low*, *medium*, and *high* as fuzzy sets to denote the relevance of the queried geometric properties (Fig. 7.6). For the inference part, we chose the *min* function for the implication, the *max* function for the aggregation, and the *centroid* function for the defuzzification. The query for identifying change points in the cycle is then modeled through the following rules:

a) IF (*average* is not *zero*), THEN (*score* is *high*)

b) IF (*average* is *zero*), THEN (*score* is *low*).

These two basic rules use the average of each segment to determine the degree of change within the cycle. The output of the FIS (bottom of Fig. 7.4) describes a score within the $[0, 1]$ interval for each segment of the time series. High scores are produced for the values where some noise was added (segments 3 to 11, and segments 16 to 30), which confirms the expected results.

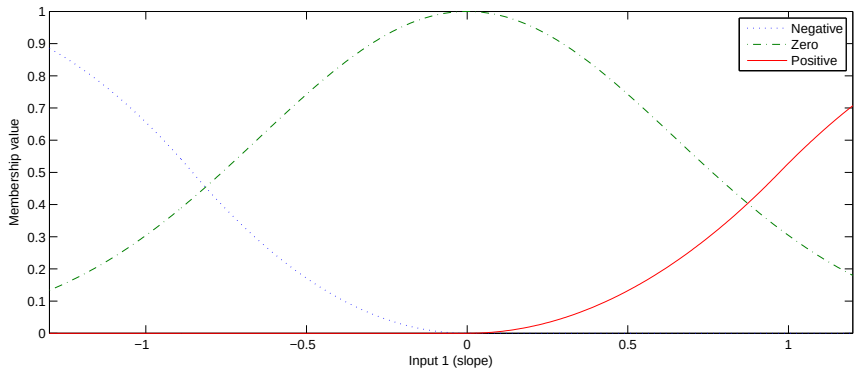


Figure 7.5: Input membership functions of the cycle time series.

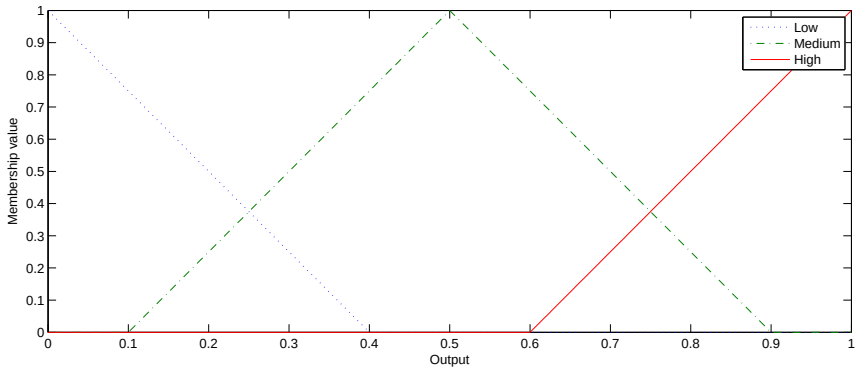


Figure 7.6: Output membership functions of the cycle time series.

7.5.2 Case study: crime trends monitoring

The objective of crime trends monitoring is to automatically detect change points within the development of crime. This case study shows how crime analysts can monitor sudden changes in the number of crimes, allowing them better to allocate resources (e.g., by sending dedicated patrols when a rise is detected). We want to emphasize that the proposed method is used on-line, meaning that we do not need the entire time series to perform these analyses and the results only depend on past values.

For this purpose, we illustrate the detection of change points with two time series from the CICOP data set. The first time series describes evening burglaries of individual houses or flats, with 72 monthly data points for a period of 6 years (top of Fig. 7.7). The second time series represents ATM break-ins, with the same sampling (top of Fig. 7.8). A more detailed description of this data set can be found in Albertetti et al. (2013b). The segmentation settings are identical for both time series of the data set: values are normalized, two disjunctive thresholds are set ($th_{DPU} = 0.05$ and $th_{SSS} = 2$; only one threshold need to be exceeded to set a new segment), and K is set to 5. The input of the FIS is the coefficient *variation* between two consecutive segments ($v_{\alpha_k}^{i->i+1}$, as in Eq. 7.18) of the average coefficient, with 5 fuzzy sets (the membership functions are shown in Fig. 7.9).

The same three inference rules for both time series were given as heuristics to querying changes in the trend:

a) IF (*var_average* is *large_decrease*), THEN (*score* is *high*)

b) IF (*var_average* is *large_increase*), THEN (*score* is *high*)

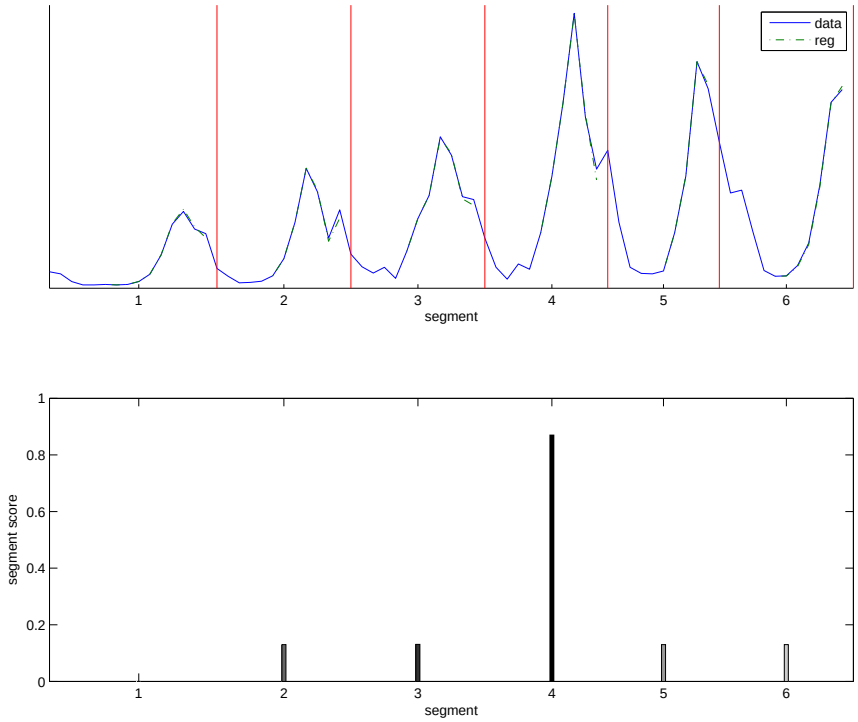


Figure 7.7: Evening burglaries time series from the CICOP data set. *(Top)* Segmentation of the time series, with 6 change points detected. *(Bottom)* Output of the FIS for each segment, representing the relevance of the query (i.e., changes in trend).

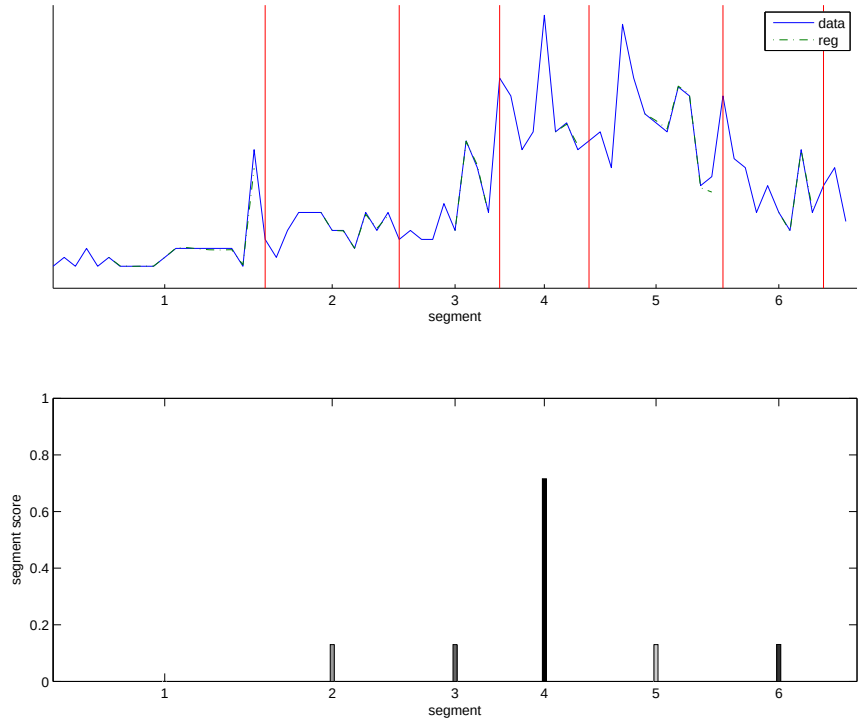


Figure 7.8: ATM break-ins time series from the CICOP data set. *(Top)* Segmentation of the time series, with 6 change points detected. *(Bottom)* Output of the FIS for each segment, representing the relevance of the query (i.e., changes in trend).

c) IF (*var_average* is *constant*), THEN (*score* is *low*).

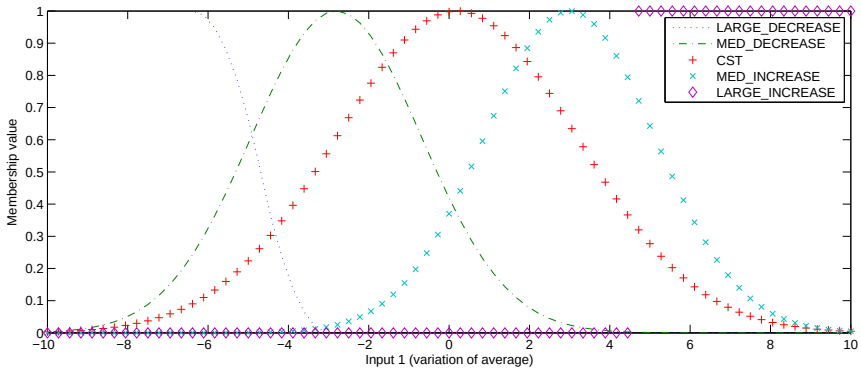


Figure 7.9: Input membership functions of both evening burglaries and ATM burglaries time series.

The highest score for both time series is observed in segment 4. Indeed, both time series are suggesting a high variation of the average after segment 3; in the next segments, the trends remain pretty stable and the score remains low.

7.5.3 Case study: change points detection on the financial TOPIX time series

To assess the general applicability of the method in a different domain, we used the TOPIX time series to detect and query change points. The weekly values consist of 522 data points (years 1985 to 1994) from the TOPIX index (*TPX:IND*, i.e., the Tokyo Stock Exchange Price Index). We also compare our results with the work of Yamanishi and Takeuchi in Yamanishi and Takeuchi (2002) and Takeuchi and Yamanishi (2006), which have used the same time series for change points and outliers detection.

The purpose of this case study is to detect change points considered as the steepest slopes of the considered time-frame. In the financial domain, it proves very useful to find change points in real-time, giving the investors a trend indicating the right moment for selling or buying stocks.

For the segmentation process, we set K to 5 and two independent thresholds ($th_{SSS} = 2$ and $th_{DPU} = 0.05$ respectively for the switches of the sign slope and the deviation of the predicted value). Figure 7.10 (top) shows the 27 change points detected.

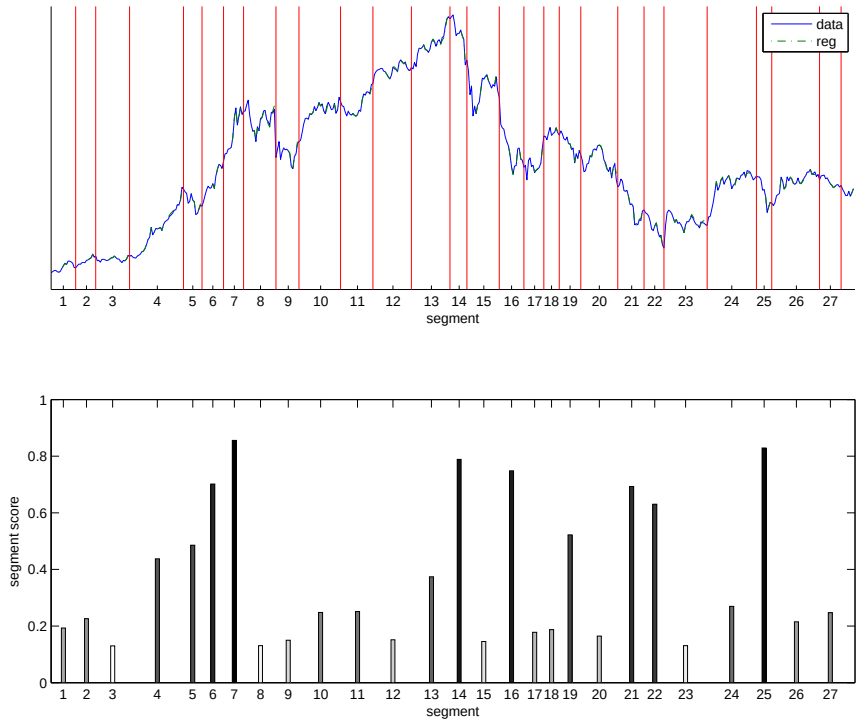


Figure 7.10: The TOPIX time series. *(Top)* Segmentation of the time series, with 27 change points detected. *(Bottom)* Output of the FIS for each segment, representing the relevance of the query (i.e., finding the steepest slopes).

The querying step is used for identifying steep slopes. As input to the FIS, the slope coefficient is used. The fuzzy sets describing the input are *negative*, *zero*, and *positive* (Fig. 7.11). The output membership is the same as for the cycle time series (Fig. 7.6). Three inference rules were given as heuristics to describe a steep slope shape:

a) IF (*slope* is *negative*), THEN (*score* is *high*)

b) IF (*slope* is *positive*), THEN (*score* is *high*)

c) IF (*slope* is *zero*), THEN (*score* is *low*).

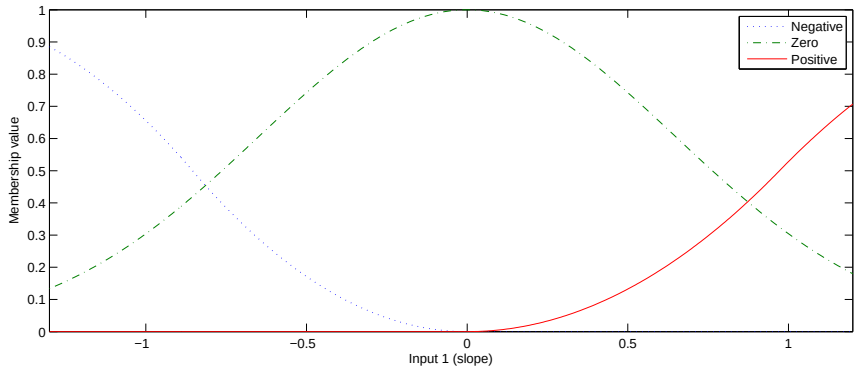


Figure 7.11: Input membership functions of the TOPIX time series.

Segments #7, #25, #14, and #16 have the highest score with FCPD (see Table 7.1). We want to emphasize that these results depend on the segmentation method, as the slope of the segment is the average of the slopes of the data points belonging to the segment.

Table 7.1: Top 4 segments in the TOPIX time series identified as the *significant changes* with the proposed method. Segment intervals are specified by values on the x-axis.

Rank	Number	Interval	Score
1	7	[112,125]	0.856
2	25	[458,468]	0.829
3	14	[259,270]	0.789
4	16	[291,307]	0.748

In their experiment (Yamanishi and Takeuchi 2002), Yamanishi and Takeuchi highlight 4 significant changes, occurring in our resulting segments #8, #2,

#14, and #25. These results are very similar to ours (i.e., segment #7 from FCPD occurs in segment #8 with Yamanishi et al., segment #24 in segment #25, segment #14 is identical, and segment #16 has no direct correspondence).

Besides, we also want to illustrate that FCPD is not limited to change points detection. Indeed, the shape space representation can be used to perform other types of analysis based on the meaningful distance computed with the shape space representation, such as clustering. In our example, we attempt to discover basic/primitive shapes in the time series. For that, we apply the K-means algorithm to the slope and curvature coefficients of the segments (i.e. the output of step 1 of the proposed method, Sect.7.4), with the number of clusters set to 4 (with the objective to delineate both negative and positive clusters of slopes and curvatures). The centroids are depicted in Fig. 7.12. The closest segments to the centroids, identified as the 4 potential primitive shapes, are shown in Fig. 7.13. One should notice that in this case the slope variable contains more information on the shape than the curvature variable.

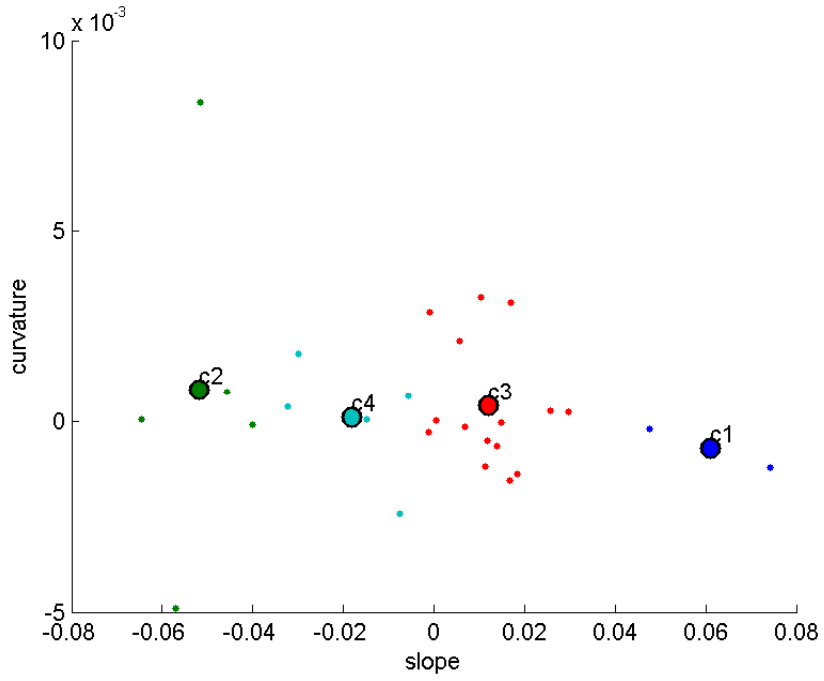


Figure 7.12: Clustering of the segments from the TOPIX time series. The segments are described with their slope and their curvature.

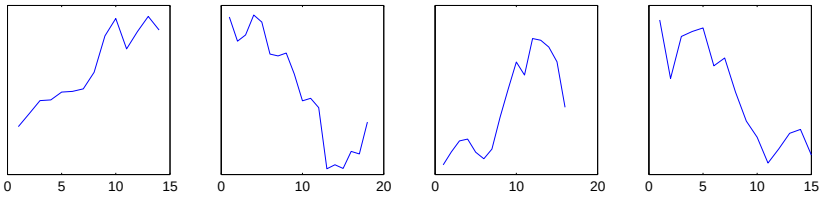


Figure 7.13: Primitive shapes of the TOPIX time series, resulting from the clusters

7.5.4 Comparison with the BFAST algorithm

Our proposed method is compared to the "break for additive seasonal and trend" (BFAST) algorithm ((Verbesselt, Hyndman, Newnham, et al. 2010) and (Verbesselt, Hyndman, Zeileis, et al. 2010)) in terms of similarity of the three most important change points detected by each method. BFAST has originally been developed for detecting changes in phenology, more precisely for climatic variations from remote sensor data. The method is however not specific to a particular data type. It uses the seasonal-trend decomposition procedure based on Loess (STL) and an estimation of breaks based on Bai (1994) with least-squares. It basically performs two steps, each being off-line, requiring access to past and future values in the computation window: first, the seasonal component is computed and removed from the observed data and second, the breakpoints are estimated.

We used the implementation of the R package *bfast* with standard parameters ($h = 0.15$, $max.iter = 1$, $season =$ "$harmonic$", $breaks = 3$) to find a maximum of 3 significant change points in both the CICOP and the SWX data sets.

The settings in FCPD are the same for both data sets, as the nature of these two data sets are similar and the objectives are identical. We set a threshold for the deviation of the predicted value (th_{DPU}) of 0.11 and K to 5. Fig. 7.14 depicts the uniform membership functions for the input variables and Fig. 7.15 for the output variable. The rules are simply relating the degree of variation to the degree of change, considering both the average and the slope:

a) IF (*var_average* or *var_slope* is *very_large_decrease*), THEN (*score* is *very_high*)

b) IF (*var_average* or *var_slope* is *large_decrease*), THEN (*score* is *high*)

c) IF (*var_average* or *var_slope* is *medium_decrease*), THEN (*score* is *medium*)

d) IF (*var_average* or *var_slope* is *small_decrease*), THEN (*score* is *low*)

e) IF (*var_average* or *var_slope* is *constant*), THEN (*score* is *very_low*)

f) IF (*var_average* or *var_slope* is *small_increase*), THEN (*score* is *low*)

g) IF (*var_average* or *var_slope* is *medium_increase*), THEN (*score* is *medium*)

h) IF (*var_average* or *var_slope* is *large_increase*), THEN (*score* is *high*)

i) IF (*var_average* or *var_slope* is *very_large_increase*), THEN (*score* is *very_high*)

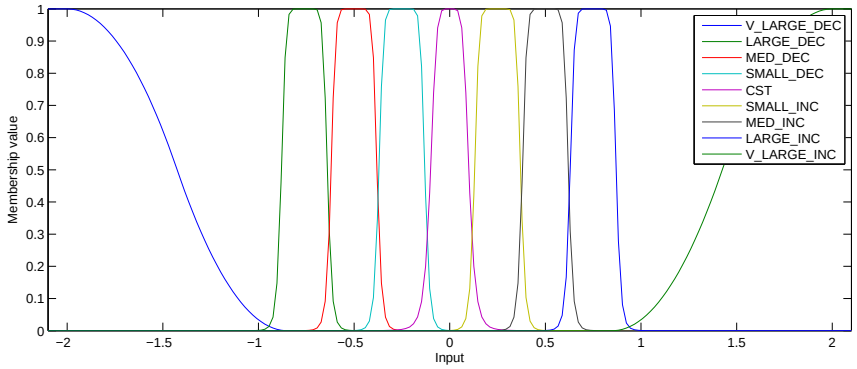


Figure 7.14: Input membership functions for the FIS of the CICOP and SWX data set.

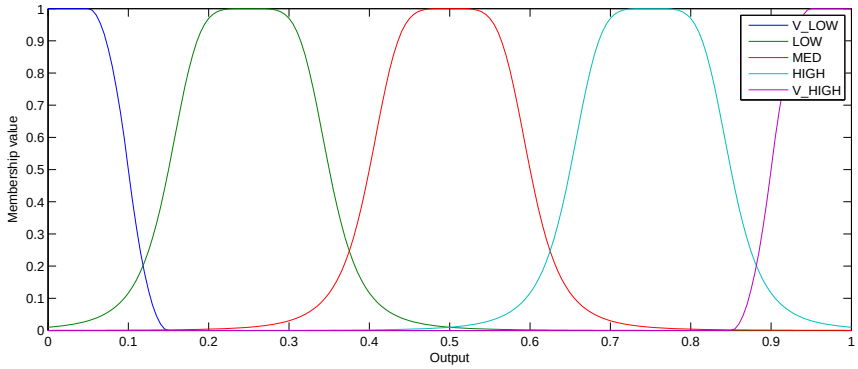
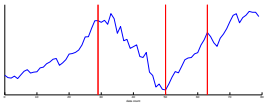


Figure 7.15: Output membership functions for the FIS of the CICOP and SWX data set.

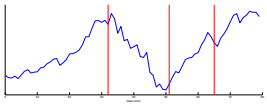
For the comparison, the 3 most significant points detected by BFAST are compared to the 3 most significant points detected by FCPD, and the offset

in the occurrence of the detection (measured in number of data points from the time domain) is reported. The offsets are the difference between each significant points from BFAST and FCPD (as illustrated in Fig. 7.16).

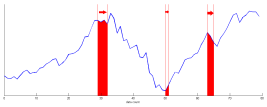
Figure 7.16: Illustration of the comparison of the 3 most significant changes between the BFAST algorithm and the proposed method.



(a) The 3 most important change points detected by BFAST.



(b) The 3 most important change points detected by FCPD.



(c) Calculation of the occurrence offset (red areas, from left to right): the 1st offset is reported as +3 data points in the x-axis for FCPD, the 2nd as +1, and the 3rd as +2.

For the 32 time series of the CICOP data set (the data set actually contains 60 time series, but we only used the time series in which at least one change was detected by BFAST for a consistent comparison), the offsets are shown in Fig. 7.17 (left).

For the 64 time series of the SWX data set (the data set consists of 70 time series, but we only considered time series where all values were defined for the concerned period), the offsets are shown in Fig. 7.17 (right).

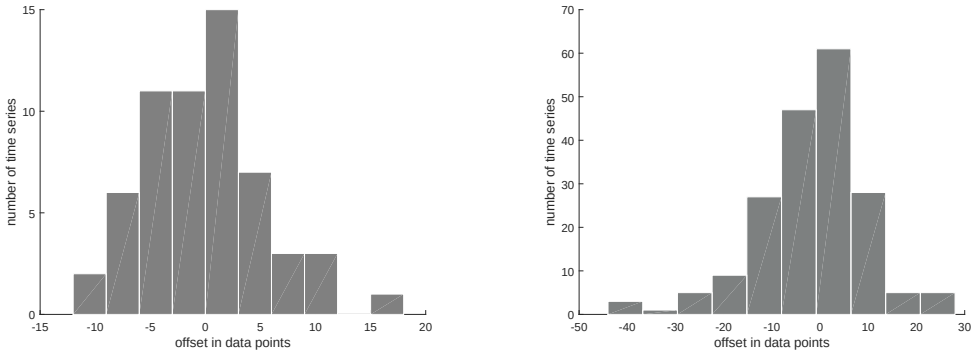
Table 7.2 summarizes this comparison. It shows that the average offset for the CICOP data set is about 4.02 data points, and about 8.15 for the SWX data set.

Table 7.2: Offset statistics for the 3 most significant changes (O1, O2, O3), compared with BFAST, in absolute values.

	O1	O2	O3	μ	σ
CICOP	4.66	3.89	3.50	**4.02**	0.48
SWX	7.09	5.88	11.49	**8.15**	2.41

The average observed absolute offsets of 4 data points for CICOP and 8 data points for SWX with BFAST are considered as pretty good results in terms of similarity, especially when we know that FCPD is on-line and therefore only past values are used to detect change points, which is not the case with

Figure 7.17: Distribution (histogram) of the offsets of the 3 most significant segments of the CICOP and the SWX data sets. A negative offset indicates that the FCPD algorithm detected the point before the BFAST algorithm.



(a) Distribution of the CICOP data set. (b) Distribution of the SWX data set.

BFAST. Because of that, in the context of crime trends monitoring, BFAST cannot be used in a real environment. The distributions of the offsets from Figure 7.17 show that very few time series present an offset exceeding the absolute value of 10 in data points, and also suggest that the offsets are equally distributed in terms of lag or of lead.

As an attempt to mimic the BFAST behavior and for comparison only, we used an automated method to fine-tune the FIS parameters, that is the settings of the membership functions, the linguistic variables and the rules. The MATLAB implementation of an adaptive-network-based fuzzy inference system (ANFIS, see Jang (1993)) was used with both the CICOP and the SWX data set to compare the results from Table 7.2, also with 5 membership functions for the same inputs. A single FIS was trained with both data sets. The consequent average offsets are higher than the "manual" version, that is 7.00 for CICOP and 11.72 for SWX.

This difficulty to extract more suitable parameters can have multiple causes. First, selecting the target of the ANFIS method (i.e., the supervised observations for the learning part) is far from obvious because many ways can be used to defined it. In our experiment, we decided to encode the score "1" when the discovered segment was within a region of ± 2 data points of a BFAST detected point, and "0" otherwise. Second, both data set might not have enough observations for ANFIS to accurately learn the parameters from a machine learning perspective. And last, most of the ANFIS implementations only support Takagi-Sugeno inference types, as a result having less flexibility

in the parameters and a different defuzzification method.

Besides, FCPD presents a huge advantage in terms of complexity. For comparison only, the most computational consuming step of BFAST, that is, the detection of breaks based on Bai (1994), is of $O(N^2)$; whereas in FCPD, for the regression, the complexity is of $O(K^2)$, where N is the number of observation and K the degree of the regression ($K \ll N$). As illustration, the running time for the SWX data set is 42 seconds for BFAST and 4 seconds for FCPD on the same computer.

7.5.5 Sensitivity analysis

In this part the sensitivity of the proposed method in regard to its parameters is evaluated. The variation of the score of a query are observed with regard to changes in the parameters of the segmentation step (i.e., K, the degree; th_{DPU} and th_{SSS}, the thresholds) and the querying step (i.e., the membership functions of the inputs). Default parameters are the ones used in the crime trends monitoring case study (Subsection 7.5.2). These variations are measured by computing the mean upper bound (i.e., the mean of the best 3 scores of the data set), the mean lower bound (i.e., the mean of the worst 3 scores of the data set), and the mean number of segments on both the CICOP and the SWX data sets.

These singular changes are introduced either on the parameter K, either on the threshold th_{DPU}, either on the threshold th_{SSS}, either in the introduction of an offset in the x-axis, or in the input membership functions (Fig. 7.18).

The membership functions are depicted in Fig. 7.19. For membership functions *TYPE_3* and *TYPE_4*, the rules have been consequently adapted (the number of membership having decreased, they need to be adapted according to the output variables):

a) IF (*var_average* or *var_slope* is *large_decrease*), THEN (*score* is *very_high*)

b) IF (*var_average* or *var_slope* is *small_decrease*), THEN (*score* is *medium*)

c) IF (*var_average* or *var_slope* is *constant*), THEN (*score* is *very_low*)

d) IF (*var_average* or *var_slope* is *small_increase*), THEN (*score* is *medium*)

e) IF (*var_average* or *var_slope* is *large_increase*), THEN (*score* is *very_high*)

To better understand these results, let us take an example with a change of the degree on the CICOP data set (top-left of Fig.7.18). The reference value,

Figure 7.18: (From top to bottom) Effects of a change in the parameter K, th_{DPU}, th_{SSS}; on the introduction of an offset ($OFFSET$); and on the input membership functions of the query (MF_TYPE). The effects are measured on the mean lower bound and the mean upper bound of the CICOP data set (left-hand) and the SWX data set (right-hand). Lower bounds and upper bounds are calculated as the average of the 3 worst/best scores. Score values are the mean of all time series of the data set. The blue regions ($K = 5$, $th_{DPU} = 0.05$, $th_{SSS} = 2$, $OFFSET = 0$, and $MF\_TYPE = 1$) are the reference values for the comparisons.

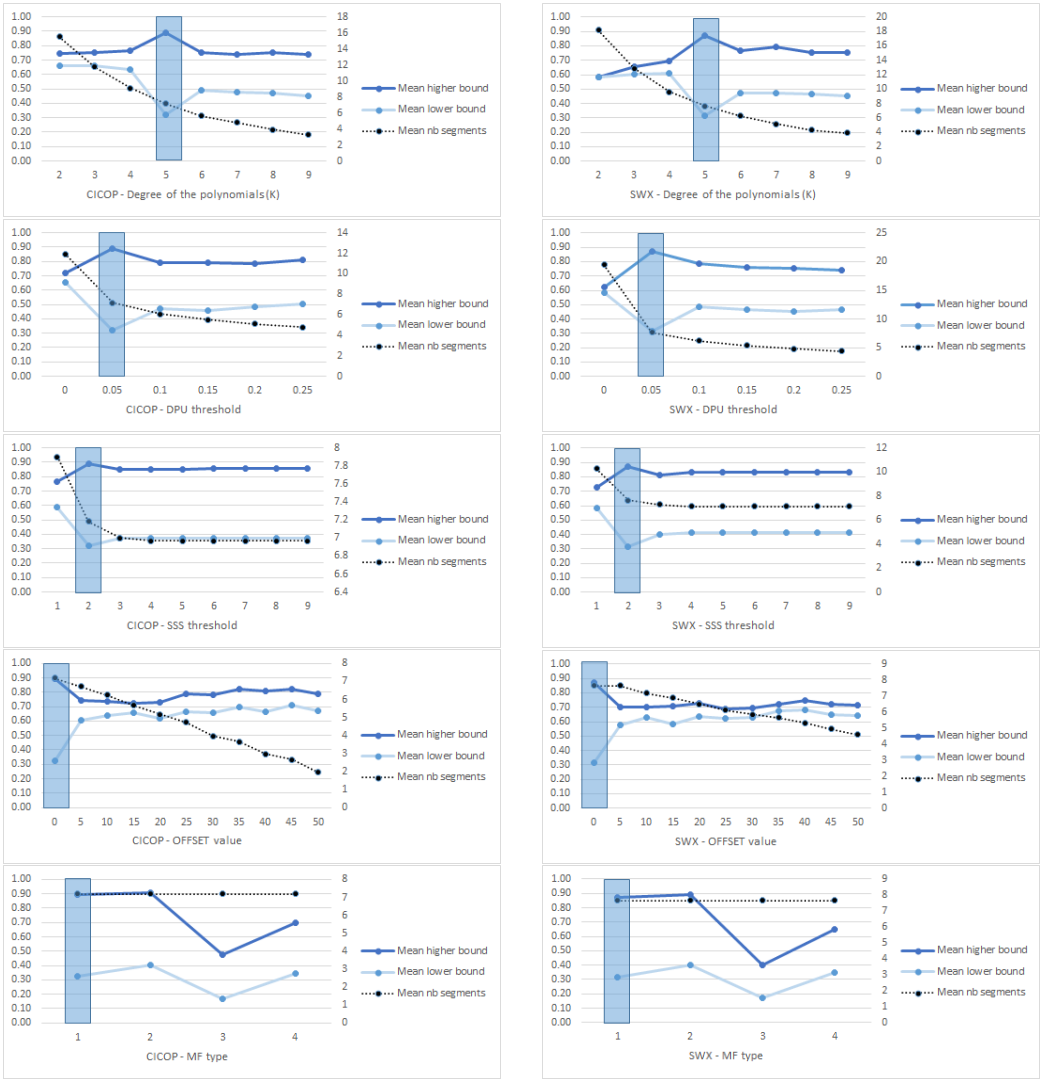
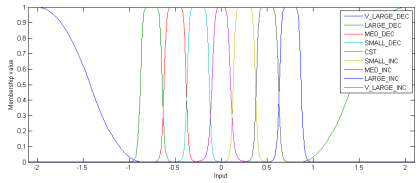
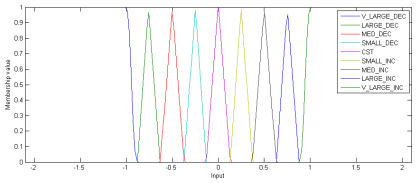


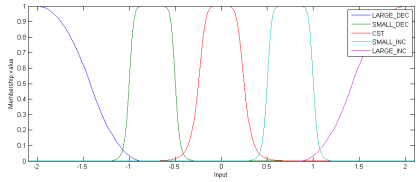
Figure 7.19: Input membership functions used for the sensitivity analysis.



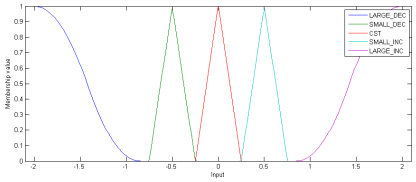
(a) Set of membership functions denoted as *TYPE_1*



(b) Set of membership functions denoted as *TYPE_2*

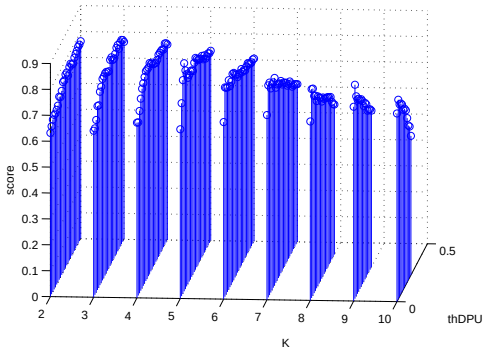


(c) Set of membership functions denoted as *TYPE_3*

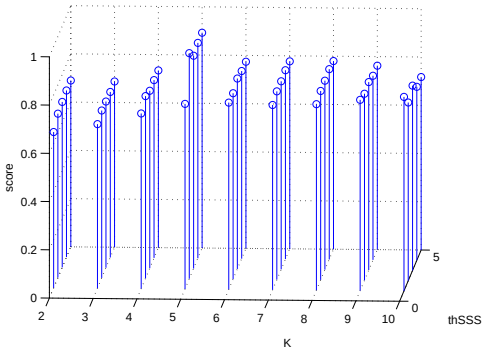


(d) Set of membership functions denoted as *TYPE_4*

Figure 7.20: Interdependence between K and the thresholds $thDPU$ and $thSSS$. The scores are the mean for both the CICOP and the SWX data sets of the top 1 segment. Missing values indicate that no segmentation was found.



(a) Interdependence between K and $thDPU$.



(b) Interdependence between K and $thSSS$.

denoted by the blue region, is shown as $K = 5$, meaning that 3 best/worst scores are defined as *reference* segments. Then, by modifying the value of K only, the score of these 6 reference segments will be compared with their mean lower bounds and mean upper bound. More generally, we can interpret these measures by saying that the bigger the difference between the lower and upper bound is, the higher the method is sensitive to the considered parameter. The effect on the mean number of segments should also be taken into account.

The first interesting observation is that the method does not seem particularly sensitive to the change of a singular effect. Indeed, the difference between the mean lower and mean upper bounds are relatively constant in most settings and for both data sets. We however denote a slightly higher difference with the thresholds.

Second, if we consider the mean upper bound, it remains high under most conditions, excepted for changes in the input membership functions. However, the mean lower bound seem to be pretty high. This could be explained by the mean number of segments, when it comes close to 6, i.e., the total number of change points considered only for the upper and lower bound.

Besides these singular changes, let us consider the interdependence of the parameters, that is between the segmentation step and the query step. The parameter K and the segmentation thresholds impact the segmentation results. If more segments are considered (i.e., by decreasing K or the thresholds, as seen in Fig. 7.18), the coefficients will describe the segments more precisely, but very local changes will be reported with respect to the query. On the other hand, settings parameters that create fewer segments (i.e., by increasing K or the thresholds), the coefficients will not be able to precisely describe the segment, resulting in inappropriate changes reported with respect to the query. Interdependence between the degree K and the thresholds $thDPU$ and $thSSS$ are shown in Fig. 7.20. The score is computed as the mean of the top 1 segment for both the CICOP and SWX data set. The variability seems to be relatively low, in the sense that modifying one parameter does not have a marked effect on the other.

7.6 Discussion

As mentioned in the previous sections, the proposed method presents three main advantages: (a) an intuitive and meaningful representation of the time

series, (b) a dynamic and on-line segmentation method, and (c) a flexible and understandable querying system.

These claims are supported by our experiments: the two case studies illustrate the flexibility and the feasibility of the intuitive querying of change points; the comparison with the BFAST algorithm show similar results in terms of accuracy; the sensibility analysis show that the parameters of the segmentation part can be consistently determined; and in terms of computational complexity, FCPD is much more efficient than BFAST.

7.7 Conclusions

A method for the detection of change points within crime-related time series was described and tested with different data sets. The combination of a meaningful representation, a dynamic segmentation, and a fuzzy inference system delivers the possibility, even for experts not related to data mining, to intuitively find change points by describing geometric properties in linguistic terms. More broadly, the considerable flexibility of the method makes possible the use of the method in any application domain, with a great potential in crime analysis.

Future work suggest further investigation on the use of mining methods to automatically discover the most appropriate membership functions of the inference system in order to mimic the behavior of existing algorithms. This alternative could present a gain in the accuracy of the detected change points, however, the opposing view is a loss in the understanding of the inference system.

Also, an implementation of a crime trends monitoring process in a real environment should be tested and the results assessed in real time by crime analysts.

Acknowledgements

The authors are grateful to the Swiss National Science Foundation (SNSF) for the support of this work under project no. 156287. The authors would also like to thank the Police de Sûreté du Canton de Vaud for the access provided to the data, and particularly Sylvain Ioset and Damien Dessimoz for their support.

"To know what you know and what you do not know, that is true knowledge."

Confucius (551 – 479 BC)

8 Conclusions

8.1 The Significance of the Contributions

The issues of knowledge extraction for crime analysis are numerous. In this dissertation, we directed the focus towards dealing with unclear objectives and uncertain environments, illustrated by three contributions.

The first contribution addressed the issue of structuring crime data. The method I proposed turns crime data from semi-structured environments, such as police reports, into well-structured and contextualized data marts. It augments and complements crime data with existing information to maximize the chances for any analytical method to extract useful knowledge.

The second contribution proposed a knowledge extraction method for crime linkage. As crime analysis deals with both qualitative and quantitative data, many existing traditional mining methods are not suitable. A fuzzy multi-criteria approach enables this possibility and also deals with uncertainties by including experts' understanding of knowledge domain.

The third contribution proposed a knowledge extraction method for change points detection in crime-related time series. The fuzzy logic layer helps crime analysts to flexibly and understandably query change points in crime-related time series.

More broadly, I investigated the extent of knowledge extraction methods for the analysis of crimes. I defined some necessary conditions to conduct analyses on crime data and reviewed some crime theories that justify the use of automated methods and guide their use. I also isolated some recurring problems emerging when crime data is analyzed and suggested some methods to deal with these difficulties.

Last but not least, this interdisciplinary research leads to several consequences, with an impact on both the computer science and the crime analysis disciplines:

- The role of knowledge extraction methods for the analysis of crime is enlightened and a deep insight is provided.
- It guides future research in both fields.
- It guides the development of real-world applications in crime analysis.
- The discovery of new crime knowledge is made easier with the use of the proposed methods. Indeed, every crime analysis unit can contribute to the advancement of knowledge with the use of their own data sets with respect to their specific environments.

8.2 Future work

To further investigate the role of knowledge extraction in crime analysis, the following research tracks are suggested.

First and foremost, more specific representations of crime data and knowledge should be considered. The efficiency of knowledge extraction methods heavily relies on the structure used to represent the data that is used to feed the algorithms.

Second, knowledge extraction methods dealing with specific tasks of crime analysis (such as link analysis or deception detection) need more support and attention from researchers. Research in public security suffers from both lack of interest and unpublished research. This latter problem often originates from the fact that results are considered as sensitive by governments and therefore outcomes are not communicated under pretext of confidentiality.

Third, a considerable effort has to be made into formalizing existing crime knowledge such as in the form of ontologies, logical rules, patterns, or processes. An open database for crime knowledge has to be built and made available to researchers. The availability of such material alone can lead to real advances in the field. Otherwise, generic algorithms without domain knowledge will continue leading contributions to using poor inference schemes and producing trivial results.

Besides, I suggest more implementations of knowledge extraction methods in real-world environments through partnerships between scientists and practitioners. Indeed, in spite of an interdisciplinary approach based on the state of the art, a well-tested method for helping crime analysts is highly valued. These latter are often confronted with the choice of mostly off-the-shelf tools provided by multinational behemoths, with very poor scientific based alternatives. Although "polishing off" a product may have little research value, it could improve the feeling crime analysts have in regard to theoretical science, considering that for the time being the role of research is underestimated within the process of creating off-the-shelf products.

After all, what is the difference between theory and practice? Between research and application? Between scientists and practitioners? I should like to finish of with this Einstein's thought:

"*In theory, theory and practice are the same. In practice, they are not.*"

Bibliography

Abraham, T. and O. de Vel (2002). "Investigative Profiling with Computer Forensic Log Data and Association Rules." In: *Proceedings of the 2002 IEEE International Conference on Data Mining*. ICDM '02. Washington, DC, USA: IEEE Computer Society, pp. 11–.

Adderley, R. W. (2007). "The Use of Data Mining Techniques in Crime Trend Analysis and Offender Profiling." PhD thesis. University of Wolverhampton.

Albertetti, F., L. Grossrieder, et al. (2016). "Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation." In: *Applied Soft Computing* 40, pp. 441–454. DOI: http://dx.doi.org/10.1016/j.asoc.2015.12.004.

Albertetti, F. and K. Stoffel (2012). "From Police Reports to Datamarts: Towards a Crime Analysis Framework." In: *Proceedings of the 5th International Workshop, IWCF 2012, Tsukuba, Japan*. Ed. by S. Srihari and K. Franke, pp. 48–59.

Albertetti, F. and K. Stoffel (2013). "An Intelligent Process-driven Knowledge Extraction Framework for Crime Analysis." In: *Abstracts Collection on New Challenges in the European Area: International Baku Forum of Young Scientists*, pp. 96–99.

Albertetti, F. et al. (2013a). "Crime linkage: A fuzzy MCDM approach." In: *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pp. 1–3. DOI: 10.1109/ISI.2013.6578772.

Albertetti, F. et al. (2013b). "The CriLiM Methodology: Crime Linkage with a Fuzzy MCDM Approach." In: *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pp. 67–74. DOI: 10.1109/EISIC.2013.17.

Atabakhsh, H. et al. (2001). *COPLINK Knowledge Management for Law Enforcement: Text Analysis, Visualization and Collaboration.*

Bai, J. (1994). "Least squares estimation of a shift in linear processes." In: *Journal of Time Series Analysis* 15.5, pp. 453–472.

Bao, Y. and L. Zhang (2010). "Decision support system based on data warehouse." In: *Journal of World Academy of Science, Engineering and Technology, 71, 172* 176.

Basseville, M., I. V. Nikiforov, et al. (1993). *Detection of abrupt changes: theory and application.* Vol. 104. Prentice Hall Englewood Cliffs.

Bennell, C. and D. Canter (2002). "Linking commercial burglaries by modus operandi: tests using regression and ROC analysis." In: *Science & Justice* 42.3, pp. 153–164. DOI: 10.1016/S1355-0306(02)71820-0.

Bennell, C., B. Snook, et al. (2012). "Computerized Crime Linkage Systems: A Critical Review and Research Agenda." In: *Criminal Justice and Behavior* 39.5, pp. 620–634. DOI: 10.1177/0093854811435210.

Berry, M. and G. Linoff (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.

Boba, R. (2009). *Crime Analysis with Crime Mapping.* Thousand Oaks, CA: Sage.

Brantingham, P. L. and P. J. Brantingham (1990). "Situational crime prevention in practice." In: *Canadian J. Criminology* 32, p. 17.

Brantingham, P. and P. Brantingham (2008). "Crime pattern theory." In: *Environmental criminology and crime analysis*, p. 78.

Brown, D. E. (1998). "The regional crime analysis program (RECAP): a framework for mining data to catch criminals." In: *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on.* Vol. 3. IEEE, pp. 2848–2853.

Canter, D. et al. (2000). "Predicting serial killers' home base using a decision support system." In: *Journal of Quantitative Criminology* 16.4, pp. 457–478.

Cao, L. (2008). "Domain Driven Data Mining (D3M)." In: *Proc. IEEE Int. Conf. Data Mining Workshops ICDMW '08*, pp. 74–76. DOI: 10.1109/ICDMW.2008.98.

Cao, L. (2010). "Domain-Driven Data Mining: Challenges and Prospects." In: *Knowledge and Data Engineering, IEEE Transactions on* 22.6, pp. 755–769.

Cao, L., R. Schurmann, and C. Thang (2005). "Domain-Driven In-Depth Pattern Discovery: A Practical Methodology." In: *Proceedings of The Australian Data Mining Conference*.

Cao, L. and C. Zhang (2007). "The Evolution Of Kdd: Towards Domain-driven Data Mining." In: *International Journal of Pattern Recognition & Artificial Intelligence* 21.4, pp. 677–692.

Cao, L., C. Zhang, and J. Liu (2006). "Ontology-based integration of business intelligence." In: *Web Intelligence and Agent Systems* 4.3, pp. 313–325.

Cao, L., H. Zhang, et al. (2011). "Combined Mining: Discovering Informative Knowledge in Complex Data." In: *Systems, Man and Cybernetics, IEEE Transactions on* 41.3, pp. 699–712.

Cappelli, C., P. D'Urso, and F. D. Iorio (2013). "Change point analysis of imprecise time series." In: *Fuzzy Sets and Systems* 225. Theme: Fuzzy Systems, pp. 23–38. DOI: 10.1016/j.fss.2013.03.001.

Carney, M. and M. Rogers (2004). "The Trojan made me do it: A first step in statistical based computer forensics event reconstruction." In: *International Journal of Digital Evidence* 2.4, pp. 1–11.

Castellano, P. and S. Sridharan (1996). "A two stage fuzzy decision classifier for speaker identification." English. In: *Speech Communication* 18.2, pp. 139–149. DOI: 10.1016/0167-6393(95)00041-0.

Cavnar, W. and J. M. Trenkle (1994). "N-Gram-Based Text Categorization." In: *Proceedings of the Thin! Annual Symposium on Document and Information Retrieval*.

Chau, M., J. J. Xu, and H. Chen (2002). *Extracting Meaningful Entities from Police Narrative Reports*.

Checkland, P. (2000). "Soft systems methodology: a thirty year retrospective." In: *Systems Research and Behavioral Science* 17, pp. 11–58.

Chen, C.-H., T.-P. Hong, and V. S. Tseng (2012). "Fuzzy data mining for time-series data." In: *Applied Soft Computing* 12.1, pp. 536–542.

Chen, H., W. Chung, et al. (2004). "Crime data mining: A general framework and some examples." In: *Computer* 37.4, pp. 50+. DOI: 10.1109/MC.2004.1297301.

Chen, H., J. Schroeder, et al. (2003). "COPLINK Connect: information and knowledge management for law enforcement." In: *Decision Support Systems* 34.3, pp. 271–285.

Chen, S.-M. (1996). "Forecasting enrollments based on fuzzy time series." In: *Fuzzy sets and systems* 81.3, pp. 311–319.

Chen, X. C. et al. (2013). "Contextual Time Series Change Detection." In: *SDM*. SIAM, pp. 503–511.

Chung, F.-l. et al. (2002). "Evolutionary time series segmentation for stock data mining." In: *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, pp. 83–90.

Clarke, R. and J. Eck (2005). *Crime analysis for problem solvers*.

Cohen, L. E. and M. Felson (1979). "Social Change and Crime Rate Trends: A Routine Activity Approach." English. In: *American Sociological Review* 44.4, pp.588–608.

Cornish, D. B. and R. V. G. Clarke (1996). *The reasoning criminal: rational choice perspectives on offending*.

Cornish, D. B. and R. V. G. Clarke (1975). *Residential treatment and its effects on delinquency*. HM Stationery Office.

Dua, S. and X. Du (2011). *Data Mining and Machine Learning in Cybersecurity*. Taylor & Francis.

Dunham, M. (2003). *Data Mining Introductory and Advanced Topics*. An Alan R. Apt book. Prentice Hall/Pearson Education.

Dursun, M. and E. E. Karsak (2010). "A fuzzy MCDM approach for personnel selection." In: *Expert Systems with Applications* 37.6, pp. 4324–4330.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). "The KDD process for extracting useful knowledge from volumes of data." In: *Communications of the ACM* 39.11, pp. 27–34. DOI: 10.1145/240455.240464.

Felson, M. and R. V. Clarke (1998). "Opportunity makes the thief." In: *Police Research Series, Paper* 98.

Ferguson, N. (1997). "Data Warehousing." In: *International Review of Law, Computers & Technology* 11.2, pp. 243–250. DOI: 10.1080/13600869755686.

Franke, K. and S. Srihari (2008). "Computational Forensics: An Overview." In: *IWCF 2008*. Lecture Notes in Computer Science 5158. Ed. by S. Srihari and K. Franke, pp. 1–10. DOI: 10.1007/978-3-540-85303-9-1.

Franke, K. and S. N. Srihari (2007). "Computational forensics: Towards hybrid-intelligent crime investigation." In: *Information Assurance and Security, 2007. IAS 2007. Third International Symposium on*. IEEE, pp. 383–386.

Fu, T.-c. (2011). "A review on time series data mining." In: *Engineering Applications of Artificial Intelligence* 24.1, pp. 164–181.

Fuchs, E., T. Gruber, J. Nitschke, et al. (2010). "Online segmentation of time series based on polynomial least-squares approximations." In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.12, pp. 2232–2245.

Fuchs, E., T. Gruber, H. Pree, et al. (2010). "Temporal data mining using shape space representations of time series." In: *Neurocomputing* 74.1, pp. 379–393.

Giannelli, P. C. (1998). "Forensic sciences." In: *Journal of Legal Medicine* 19.3, pp. 463–470. DOI: 10.1080/01947649809511072.

Gottlieb, S., I. A. Sheldon, and S. Raj (1994). *Crime Analysis : From First Report to Final Arrest*. Montclair, Ca.: Alpha: Print.

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2013a). "Patterns and trends detection in crime analysis: contribution of data mining techniques."

In: *13th Annual Conference of the European Society of Criminology (ESC). Budapest, Hungary.*

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2013b). "Des données aux connaissances, un chemin difficile: réflexion sur la place du data mining en analyse criminelle." In: *Revue Internationale de Criminologie et de Police Technique et Scientifique* 66, pp. 99–116.

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2014a). "Détection de tendances en renseignement criminel: contribution des techniques de data mining." In: *XIVème Colloque de l'Association Internationale des Criminologues de Langue Française (AICLF). Liège, Belgique.*

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2014b). "Trends detection in crime analysis: the contribution of computational methods." In: *22nd International Symposium on the Forensic Sciences (ANZFSS).*

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2015). "Computational Forensic Criminology framework: application on crime trends analysis." In: *15th Annual Conference of the European Society of Criminology (ESC). Budapest, Hungary.*

Grossrieder, L. et al. (2012a). "An Intelligent Process-driven Knowledge Extraction Framework for Crime Analysis." In: *12th Annual Conference of the European Society of Criminology (ESC). Bilbao, Spain.*

Grossrieder, L. et al. (2012b). "Extraction intelligente de connaissances axée sur les processus dans le cadre du renseignement criminel." In: *XIIIème Colloque de l'Association Internationale des Criminologues de Langue Française (AICLF). Montreal, Canada.*

Grubesic, T. (2006). "On the application of fuzzy clustering for crime hot spot detection." In: *Journal Of Quantitative Criminology* 22.1, pp. 77–105. DOI: 10.1007/s10940-005-9003-6.

Grubin, D. et al. (2001). *Linking serious sexual assaults through behaviour.* Home Office London.

Güler, I. and E. D. Übeyli (2005). "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients." In: *Journal of neuroscience methods* 148.2, pp. 113–121.

Güner, H. A. A. and H. A. Yumuk (2014). "Application of a fuzzy inference system for the prediction of longshore sediment transport." In: *Applied Ocean Research* 48, pp. 162–175.

Hauck, R. V. and H. Chen (1999). *COPLINK: A Case of Intelligent Analysis and Knowledge Management*.

Hazelwood, R. R. and J. I. Warren (2004). "Linkage analysis: modus operandi, ritual, and signature in serial sexual crime." In: *Aggression and Violent Behavior* 9.3, pp. 307–318. DOI: 10.1016/j.avb.2004.02.002.

Heaton, R. (2000). "The prospects for intelligence-led policing: Some Historical and quantitative considerations." In: *Policing and Society* 9.4, pp. 337–355. DOI: 10.1080/10439463.2000.9964822.

Henig, M. I. and J. T. Buchanan (1996). "Solving MCDM problems: Process concepts." In: *Journal of Multi-Criteria Decision Analysis* 5.1, pp. 3–21. DOI: 10.1002/(SICI)1099-1360(199603)5:1<3::AID-MCDA85>3.0.CO;2-6.

Hess, K., C. Orthmann, and H. Cho (2010). *Police Operations: Theory and Practice*. Cengage Learning.

Huarng, K. (2001). "Heuristic models of fuzzy time series for forecasting." In: *Fuzzy sets and systems* 123.3, pp. 369–386.

Hwang, J.-R., S.-M. Chen, and C.-H. Lee (1998). "Handling forecasting problems using fuzzy time series." In: *Fuzzy sets and systems* 100.1, pp. 217–228.

Imhoff, C., N. Galemmo, and J. Geiger (2003). *Mastering data warehouse design: relational and dimensional techniques*. Timely, practical, reliable. Wiley Pub.

Inmon, W. (2002). *Building the data warehouse*. Wiley.

International Association of Crime Analysts (IACA) (2014). *Definition and Types of Crime Analysis*. Standards, Methods, & Technology (SMT) Committee White Paper.

Jaccard, P. (1908). *Nouvelles recherches sur la distribution florale*.

Jang, J. (1993). "Anfis - Adaptive-network-based Fuzzy Inference System." In: *Systems, Man and Cybernetics, IEEE Transactions on* 23.3, pp. 665–685. DOI: 10.1109/21.256541.

Jukic, N. (2006). "Modeling Strategies And Alternatives For Data Warehousing Projects." In: *Communications of the ACM* 49.4, pp. 83–88.

Kahraman, C. (2008). *Fuzzy multi-criteria decision making: theory and applications with recent developments*. Vol. 16. Springer.

Keogh, E. J. et al. (2001). "An Online Algorithm for Segmenting Time Series." In: *ICDM*, pp. 289–296. DOI: 10.1109/ICDM.2001.989531.

Keogh, E. et al. (2004). "Segmenting time series: A survey and novel approach." In: *Data mining in time series databases* 57, pp. 1–22.

Kimball, R. (1997). *A Dimensional Modeling Manifesto*.

Kimball, R. (2004). *The Data Warehouse ETL Toolkit*. Wiley, New York.

Kind, S. S. (1994). "Crime investigation and the criminal trial: a three chapter paradigm of evidence." In: *Journal of the Forensic Science Society* 34.3, pp. 155–164. DOI: http://dx.doi.org/10.1016/S0015-7368(94)72908-X.

Kloptchenko, A. et al. (2004). "Combining data and text mining techniques for analysing financial reports." English. In: *Intelligent Systems in Accounting, Finance and Management* 12.1, pp. 29–41.

Kohn, M., J. H. P. Eloff, and M. S. Olivier (2006). "Framework for a Digital Forensic Investigation." In: *Proceedings of the ISSA 2006 from Insight to Foresight Conference*. Ed. by H. S. Venter et al. Published electronically. Sandton, South Africa.

Krohling, R. and D. Rigo (2009). "Fuzzy Group Decision Making for Management of Oil Spill Responses." In: *Applications of Soft Computing* 58, pp. 3–12.

Kuhn, A. (2000). "Le Chaos en Criminologie." In: *Revue Internationale de Criminologie et de Police Technique et Scientifique* 4. Ed. by Polymedia, pp. 404–419.

Kumar, K. and B. Wu (2001). "Detection of change points in time series analysis with fuzzy statistics." In: *International Journal of Systems Science* 32.9, pp. 1185–1192. DOI: 10.1080/00207720110034698.

Kurgan, L. A. and P. Musilek (2006). "A survey of Knowledge Discovery and Data Mining process models." In: *The Knowledge Engineering Review* 21.01, pp. 1–24. DOI: 10.1017/S0269888906000737.

Last, M., Y. Klein, and A. Kandel (2001). "Knowledge discovery in time series databases." In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 31.1, pp. 160–169.

Lee, C.-H. L., A. Liu, and W.-S. Chen (2006). "Pattern discovery of fuzzy time series for financial prediction." In: *Knowledge and Data Engineering, IEEE Transactions on* 18.5, pp. 613–625.

Lee, I. and V. Estivill-Castro (2011). "Exploration Of Massive Crime Data Sets Through Data Mining Techniques." In: *Applied Artificial Intelligence* 25.5, pp. 362–379.

Li, C. (2009). *Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions*. IGI Global.

Liao, N., S. Tian, and T. Wang (2009). "Network forensics based on fuzzy logic and expert system." English. In: *Computer Communications* 32.17, pp. 1881–1892. DOI: 10.1016/j.comcom.2009.07.013.

Lin, J., E. Keogh, S. Lonardi, et al. (2003). "A symbolic representation of time series, with implications for streaming algorithms." In: *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, pp. 2–11.

Lin, J., E. Keogh, L. Wei, et al. (2007). "Experiencing SAX: a novel symbolic representation of time series." In: *Data Mining and Knowledge Discovery* 15.2, pp. 107–144.

Maltoni, D. et al. (2009). *Handbook of Fingerprint Recognition*. Springer professional computing. Springer.

Mamdani, E. (1974). "Application of fuzzy algorithms for control of simple dynamic plant." In: *Proceedings of the Institution of Electrical Engineers* 121.12, pp. 1585–1588. DOI: 10.1049/piee.1974.0328.

Mamdani, E. and S. Assilian (1999). "An experiment in linguistic synthesis with a fuzzy logic controller." English. In: *International Journal of Human-Computer Studies* 51.2, pp. 135–147. DOI: 10.1006/ijhc.1973.0303.

Maxfield, M. G. (1999). "The National Incident-Based Reporting System: Research and Policy Applications." In: *Journal of Quantitative Criminology* 15 (2), pp. 119–149. DOI: 10.1023/A:1007518620521.

Melnyk, T. et al. (2011). "Another look at across-crime similarity coefficients for use in behavioural linkage analysis: an attempt to replicate Woodhams, Grant, and Price (2007)." In: *Psychology, Crime & Law* 17.4, pp. 359–380. DOI: 10.1080/10683160903273188.

Mena, J. (2011). *Machine Learning Forensics for Law Enforcement, Security, and Intelligence*. Taylor & Francis.

Mikut, R. and M. Reischl (2011). "Data mining tools." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.5, pp. 431–443. DOI: 10.1002/widm.24.

Moody, D. L. and M. A. Kortink (2000). "From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design." In: *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000)*, pp. 5–16.

Morelato, M. et al. (2014). "Forensic intelligence framework-Part I: Induction of a transversal model by comparing illicit drugs and false identity documents monitoring." In: *Forensic science international* 236, pp. 181–190.

Moreno-Garcia, J. et al. (2014). "The generation of qualitative descriptions of multivariate time series using fuzzy logic." In: *Applied Soft Computing* 23, pp. 546–555. DOI: http://dx.doi.org/10.1016/j.asoc.2014.05.021.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press.

Oatley, G., J. Zeleznikow, and B. Ewart (2005). "Matching and Predicting Crimes." English. In: *Applications and Innovations in Intelligent Systems XII*. Ed. by A. Macintosh, R. Ellis, and T. Allen. Springer London, pp. 19–32. DOI: 10.1007/1-84628-103-2-2.

Olivier, M. S. (2009). "On metadata context in Database Forensics." In: *Digital Investigation* 5.3–4, pp. 115–123. DOI: 10.1016/j.diin.2008.10.001.

Plasse, M. et al. (2007). "Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set." In: *Computational Statistics & Data Analysis* 52.1, pp. 596–613. DOI: 10.1016/j.csda.2007.02.020.

Ratcliffe, J. (2008). *Intelligence-Led Policing*. Willan.

Ratcliffe, J. (2009). *Strategic Thinking in Criminal Intelligence*. Federation Press.

Ray, I. and S. Shenoi (2008). *Advances in Digital Forensics IV*. International Federation for Information Processing. Springer.

Reeves, J. et al. (2007). "A review and comparison of changepoint detection techniques for climate data." In: *Journal of Applied Meteorology and Climatology* 46.6, pp. 900–915.

Ribaux, O., A. Girod, et al. (2003). "Forensic intelligence and crime analysis." In: *Law, Probability and Risk* 2.1, pp. 47–60.

Ribaux, O. and P. Margot (1999). "Inference structures for crime analysis and intelligence: the example of burglary using forensic science data." In: *Forensic Science International* 100.3, pp. 193–210. DOI: 10.1016/S0379-0738(98)00213-8.

Ribaux, O., A. Baylon, E. Lock, et al. (2010). "Intelligence-led crime scene processing. Part II: Intelligence and crime scene examination." In: *Forensic Science International* 199.1-3, pp. 63–71.

Ribaux, O., A. Baylon, C. Roux, et al. (2010). "Intelligence-led crime scene processing. Part I: Forensic intelligence." In: *Forensic Science International* 195.1-3, pp. 10–16.

Ribaux, O., T. Genessay, and P. Margot (2011). "Les processus de veille opérationnelle et science forensique." In: *Sphères de surveillance*. Ed. by S. Leman-Langlois. Les Presses de l'Université de Montréal. Montréal, pp. 137–158.

Ribaux, O., S. J. Walsh, and P. Margot (2006). "The contribution of forensic science to crime analysis and investigation: forensic intelligence." In: *Forensic Science International* 156.2-3, pp. 171–181.

Rossmo, D. K. (2000). *Geographic profiling*. CRC press.

Rossy, Q. et al. (2013). "Integrating forensic information in a crime intelligence database." In: *Forensic Science International*. DOI: 10.1016/j.forsciint.2012.10.010.

Santtila, P. et al. (2008). "Behavioural crime linking in serial homicide." In: *Psychology, Crime & Law* 14.3, pp. 245–265. DOI: 10.1080/10683160701739679.

Schroeder, J. and T. P. Dept (2001). *COPLINK: Database Integration and Access for a Law Enforcement Intranet, Final Report*.

Shepherd, J. (2004). "What Is the Digital Era?" In: *Social and Economic Transformation in the Digital*. Print, pp. 1–18.

Sherman, L. W., P. R. Gartin, and M. E. Buerger (1989). "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place." In: *Criminology* 27.1, pp. 27–56.

Sinha, P. (1998). "A symmetry perceiving adaptive neural network and facial image recognition." English. In: *Forensic Science International* 98.1-2, pp. 67–89. DOI: 10.1016/S0379-0738(98)00137-6.

Snook, B. et al. (2012). "The Violent Crime Linkage Analysis System A Test of Interrater Reliability." In: *Criminal Justice and Behavior* 39.5, pp. 607–619.

Song, Q. and B. S. Chissom (1993a). "Forecasting enrollments with fuzzy time series—part I." In: *Fuzzy sets and systems* 54.1, pp. 1–9.

Song, Q. and B. S. Chissom (1993b). "Fuzzy time series and its models." In: *Fuzzy sets and systems* 54.3, pp. 269–277.

Song, Q. and N. Kasabov (2000). "Dynamic evolving neuro-fuzzy inference system (DENFIS): On-line learning and application for time-series prediction." In: *Proc. 6th International Conference on Soft Computing*. Citeseer, pp. 696–701.

Steen, M. van der and M. Blom (2007). *A roadmap for future forensic research*. Tech. rep. Technical report, Netherlands Forensic InsTtute (NFI), The Hague, The Netherlands.

Stoffel, K., D. Han, and P. Cotofrei (2010). "Fuzzy Methods for Forensic Data Analysis." In: *Proceedings of International Conference SoCPaR 2010*.

Stoffel, K. and P. Cotofrei (2011). "Fuzzy Extended BPMN for Modelling Crime Analysis Processes." In: *SIMPDA 2011*.

Stoffel, K., P. Cotofrei, and D. Han (2011). "Fuzzy Clustering based Methodology for Multidimensional Data Analysis in Computational Forensic Domain." In: *ITCF 2010*.

Straccia, U. (2009). "Multi Criteria Decision Making in Fuzzy Description Logics: A First Step." In: *KES (1)*, pp. 78–86. DOI: 10.1007/978-3-642-04595-0-10.

Sugeno, M. and T. Takagi (1985). "Fuzzy Identification of Systems and its Applications to Modeling and Control." English. In: *IEEE Transactions On Systems Man And Cybernetics* 15.1, pp. 116–132.

Takeuchi, J.-i. and K. Yamanishi (2006). "A unifying framework for detecting outliers and change points from time series." In: *Knowledge and Data Engineering, IEEE Transactions on* 18.4, pp. 482–492.

Tonkin, M., T. Grant, and J. W. Bond (2008). "To link or not to link: a test of the case linkage principles using serial car theft data." In: *Journal of Investigative Psychology and Offender Profiling* 5.1-2, pp. 59–77. DOI: 10.1002/jip.74.

Vel, O. de et al. (2001). "Mining e-mail content for author identification forensics." English. In: *Sigmod Record* 30.4, pp. 55–64.

Verbesselt, J., R. Hyndman, G. Newnham, et al. (2010). "Detecting trend and seasonal changes in satellite image time series." In: *Remote sensing of Environment* 114.1, pp. 106–115.

Verbesselt, J., R. Hyndman, A. Zeileis, et al. (2010). "Phenological change detection while accounting for abrupt and gradual trends in satellite image time series." In: *Remote Sensing of Environment* 114.12, pp. 2970–2980.

Wache, H. et al. (2001). *Ontology-Based Integration of Information - A Survey of Existing Approaches*.

Wallenius, J. et al. (2008). "Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead." In: *Management Science* 54.7, pp. 1336–1349.

Wang, G., H. Chen, and H. Atabakhsh (2004). "Automatically detecting deceptive criminal identities." In: *Communications of the ACM* 47.3, pp. 70–76.

Wang, H., D. Zhang, and K. G. Shin (2004). "Change-Point Monitoring for the Detection of DoS Attacks." English. In: *IEEE Transactions on Dependable and Secure Computing* 1.4. Copyright - Copyright IEEE Computer Society Oct-Dec 2004; Caractéristique du document - references; Dernière mise à jour - 2010-06-07, pp. 193–208.

Warren, J. et al. (1998). "Crime Scene and Distance Correlates of Serial Rape." English. In: *Journal of Quantitative Criminology* 14 (1), pp. 35–59. DOI: 10.1023/A:1023044408529.

Weisel, D. L. (2005). *Analyzing repeat victimization*. US Department of Justice, Office of Community Oriented Policing Services.

Westphal, C. (2008). *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. Taylor & Francis.

Wilson, T. F. and P. L. Woodard (1987). *Automated fingerprint identification systems: Technology and policy issues*. US Department of Justice, Bureau of Justice Statistics.

Witten, I., E. Frank, and M. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Wolfgang, M. E., R. Figlio, and T. Sellin (1987). *Delinquency in a Birth Cohort*. Studies in Crime and Justice. University of Chicago Press.

Wu, B. and M.-H. Chen (1999). "Use of fuzzy statistical technique in change periods detection of nonlinear time series." In: *Applied Mathematics and Computation* 99.2, pp. 241–254.

Wu, C.-A. et al. (2011). "Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining." In: *Expert Systems with Applications* 38.9, pp. 11011–11023. DOI: 10.1016/j.eswa.2011.02.144.

Xu, J. and H. Chen (2004). "Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks." English. In: *Decision Support Systems* 38.3, pp. 473–487. DOI: 10.1016/S0167-9236(03)00117-9.

Yamanishi, K. and J.-i. Takeuchi (2002). "A unifying framework for detecting outliers and change points from non-stationary time series data." In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 676–681.

Yu, J.-R., G.-H. Tzeng, and H.-L. Li (2001). "General fuzzy piecewise regression analysis with automatic change-point detection." In: *Fuzzy Sets and Systems* 119.2, pp. 247–257. DOI: http://dx.doi.org/10.1016/S0165-0114(98)00384-4.

Zadeh, L. (1965). "Fuzzy sets." In: *Information and Control* 8.3, pp. 338–353. DOI: 10.1016/S0019-9958(65)90241-X.

Zhao, Y. et al. (2008). "Combined Pattern Mining: From Learned Rules to Actionable Knowledge." In: *Proceedings of the 21th Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*. AI '08. Auckland, New Zealand: Springer-Verlag, pp. 393–403. DOI: 10.1007/978-3-540-89378-3-40.

Zounemat-Kermani, M. and M. Teshnehlab (2008). "Using adaptive neuro-fuzzy inference system for hydrological time series prediction." In: *Applied Soft Computing* 8.2, pp. 928–936.

Author's Publications

Albertetti, F., L. Grossrieder, et al. (2016). "Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation." In: *Applied Soft Computing* 40, pp. 441–454. DOI: http://dx.doi.org/10.1016/j.asoc.2015.12.004.

Albertetti, F. and K. Stoffel (2012). "From Police Reports to Datamarts: Towards a Crime Analysis Framework." In: *Proceedings of the 5th International Workshop, IWCF 2012, Tsukuba, Japan*. Ed. by S. Srihari and K. Franke, pp. 48–59.

Albertetti, F. and K. Stoffel (2013). "An Intelligent Process-driven Knowledge Extraction Framework for Crime Analysis." In: *Abstracts Collection on New Challenges in the European Area: International Baku Forum of Young Scientists*, pp. 96–99.

Albertetti, F. et al. (2013a). "Crime linkage: A fuzzy MCDM approach." In: *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pp. 1–3. DOI: 10.1109/ISI.2013.6578772.

Albertetti, F. et al. (2013b). "The CriLiM Methodology: Crime Linkage with a Fuzzy MCDM Approach." In: *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pp. 67–74. DOI: 10.1109/EISIC.2013.17.

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2013a). "Patterns and trends detection in crime analysis: contribution of data mining techniques." In: *13th Annual Conference of the European Society of Criminology (ESC). Budapest, Hungary*.

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2013b). "Des données aux connaissances, un chemin difficile: réflexion sur la place du data mining

en analyse criminelle." In: *Revue Internationale de Criminologie et de Police Technique et Scientifique* 66, pp. 99–116.

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2014a). "Détection de tendances en renseignement criminel: contribution des techniques de data mining." In: *XIVème Colloque de l'Association Internationale des Criminologues de Langue Française (AICLF). Liège, Belgique.*

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2014b). "Trends detection in crime analysis: the contribution of computational methods." In: *22nd International Symposium on the Forensic Sciences (ANZFSS).*

Grossrieder, L., F. Albertetti, K. Stoffel, and O. Ribaux (2015). "Computational Forensic Criminology framework: application on crime trends analysis." In: *15th Annual Conference of the European Society of Criminology (ESC). Budapest, Hungary.*

Grossrieder, L. et al. (2012a). "An Intelligent Process-driven Knowledge Extraction Framework for Crime Analysis." In: *12th Annual Conference of the European Society of Criminology (ESC). Bilbao, Spain.*

Grossrieder, L. et al. (2012b). "Extraction intelligente de connaissances axée sur les processus dans le cadre du renseignement criminel." In: *XIIIème Colloque de l'Association Internationale des Criminologues de Langue Française (AICLF). Montreal, Canada.*