Des données aux connaissances, un chemin difficile: réflexion sur la place du data mining en analyse criminelle (1)

par **Lionel GROSSRIEDER***, **Fabrizio ALBERTETTI****, **Kilian STOFFEL***** et **Olivier RIBAUX******

Résumé:
Le «data mining», ou «fouille de données», est un ensemble de méthodes et de techniques attractif qui a connu une popularité fulgurante ces dernières années, spécialement dans le domaine du marketing. Le développement récent de l'analyse ou du renseignement criminel soulève des problématiques auxquelles il est tentant d'appliquer ces méthodes et techniques. Le potentiel et la place du data mining dans le contexte de l'analyse criminelle doivent être mieux définis afin de piloter son application. Cette réflexion est menée dans le cadre du renseignement produit par des systèmes de détection et de suivi systématique de la criminalité répétitive, appelés processus de veille opérationnelle. Leur fonctionnement nécessite l'existence de patterns inscrits dans les données, et justifiés par les approches situationnelles en criminologie. Muni de ce bagage théorique, l'enjeu principal revient à explorer les possibilités de détecter ces patterns au travers des méthodes et techniques de data mining. Afin de répondre à cet objectif, une recherche est actuellement menée en Suisse à travers une approche interdisciplinaire combinant des connaissances forensiques, criminologiques, et computationnelles.

Mots-clés: Analyse criminelle, Renseignement criminel, Data mining, Criminologie environnementale, Renseignement forensique

Summary:
Data mining is an attractive set of techniques which had seen its popularity grown up in recent years, especially in marketing. The recent development of crime analysis and crime intelligence reveals issues that concern data mining, and temptation is great to apply these techniques. A reflection about place of data mining in the context of crime analysis is needed to control its application. This reflection is grounded within an intelligence-led framework provide by detection and follow-up systems called operational monitoring. Their using needs existence of patterns among data, which are justified by situational approaches in criminology. With this theoretical knowledge, the main purpose is to explore the possibility to detect these patterns trough data mining techniques. To reach empirically these objectives, a research is actually led in Switzerland through an interdisciplinary approach combining forensic, criminological, and computational knowledge.

Keywords: Crime analysis, Crime intelligence, Data mining, Environmental criminology, Forensic intelligence

* Doctorant, Institut de Police Scientifique, Université de Lausanne, Suisse.
** Doctorant, Institut du Management de l'Information, Université de Neuchâtel, Suisse.
*** Professeur, Institut du Management de l'Information, Université de Neuchâtel, Suisse.
**** Professeur, Institut de Police Scientifique, Université de Lausanne, Suisse.

Introduction

Les mutations de la criminalité (e.g. mobilité accrue, développement dans de nouveaux espaces numériques et économiques) et la nouvelle traçabilité des activités humaines par l'usage des technologies de l'information et de la communication demandent de renforcer le traitement des informations et le renseignement qui pilotent l'action de sécurité. Afin de tendre vers cet objectif, le développement de l'analyse criminelle ces vingt dernières années a permis d'appréhender le phénomène criminel au-delà des considérations légales traditionnelles, apportant un support tant au niveau de l'investigation que du renseignement.

Pour détecter, analyser et comprendre des problèmes de criminalité, il est nécessaire de puiser dans diverses disciplines afin d'explorer toutes les facettes qui les composent. L'analyse ou le renseignement criminel dans son sens large, doit non seulement faire appel à des connaissances criminologiques, mais également des connaissances sur les possibilités d'exploiter les traces du crime, c.-à-d. la science forensique. De plus, la constante augmentation des données à traiter et la nécessité de célérité quant au renseignement produit rendent les techniques computationnelles, et particulièrement les techniques de data mining, attractives pour l'analyse criminelle. Cependant, le data mining suscite également une inquiétude vis-à-vis de son application dans un domaine sensible tel que l'analyse criminelle, notamment à cause des atteintes perçues ou possibles aux libertés individuelles. On comprend dès lors l'importance de maîtriser l'ensemble de la technologie envisagée. Le data mining, qui peut être traduit par «fouille de données» ou encore «forage de données», est un ensemble de méthodes et techniques dont le but est l'exploration et l'analyse de grandes banques de donnée afin de détecter des règles et des patterns inconnus ou dissimulés (Tufféry, 2007). Il est donc susceptible d'aider à détecter et comprendre les patterns imprimés par les activités criminelles dans les données qu'elles génèrent.

Dans son ambition extrême, le data mining s'exprime de manière abstraite dans un processus global qui met en perspective l'information afin de favoriser et orienter la prise de décision. Appelé KDD (***K**nowledge **D**iscovery in **D**atabases*), ce processus cherche, en partant des données, à produire des connaissances (Fayyad, Piatetsky-Shapiro et Smyth, 1996). Idéalement, cette boîte noire malaxerait les grands ensembles de données et nous dirait tout sur le phénomène criminel. Par exemple, toute la variété des formes des répétitions criminelles, ainsi que les schémas criminels prédictibles devraient se dégager d'eux-mêmes par ces traitements, sans devoir les guider par des connaissances *a priori* sur le crime. Ces progrès offrent l'espoir pour les analystes de se débarrasser des *a priori* qui renforcent leur propension à rechercher dans les données les schémas ou patterns qu'ils connaissent déjà, sans possibilité d'en découvrir de nouveaux, évitant ainsi ce que l'on appelle le tunnel mental ou «l'effet tunnel» (Kahneman, 2011).

Bien sûr, ce paradigme traditionnel du data mining, dit centré sur les données elles-mêmes, a ses limites. D'autres courants récents préconisent une

approche plus pragmatique. La mise en œuvre des techniques de data mining est orientée en alimentant le processus de connaissances et de contraintes imposées par le contexte de l'analyse. Le *Domain Driven Data Mining* (D^3M) cherche par exemple à combler ce gouffre entre les idéaux académiques et les objectifs métiers (Cao, 2008; Cao, Yu, Zhang et Zhao, 2010). Pour ce faire, le D^3M a besoin d'intégrer des connaissances à toutes les étapes du processus KDD. L'argument étant que seule une connaissance approfondie du domaine d'intérêt permet de dégager des résultats utiles et pertinents à partir des données. Sans quoi, tous ces efforts calculatoires risquent de ne produire que des résultats triviaux ou incohérents.

Un des grands défis de l'application du data mining en analyse criminelle est de gérer l'équilibre entre ces deux visions. Il faut trouver le juste milieu entre la nécessité de retirer les œillères de l'analyste en engageant ces méthodes et techniques et le pragmatisme de son application théoriquement dirigée. Afin d'approcher ce point d'équilibre, aussi subtil que fragile, il convient de replacer le data mining dans le contexte de l'analyse criminelle, de sorte à comprendre son rôle, son utilisation, et ses limites.

Des données aux patterns

Durant les vingt dernières années, l'investigation criminelle a fait face à un flux de données croissant qui rend d'autant plus difficile leur exploitation. Une des solutions proposée fut l'exploitation systématique de la recherche et de la gestion des liens (Ribaux, Taroni et Margot, 1995; Ribaux, 1997), avec comme objectif de regrouper les répétitions criminelles. L'analyse des problèmes qui s'en dégagent oriente les décisions sur les priorités, le choix des stratégies, et la définition des mesures opérationnelles capables de les traiter. Ce suivi de la criminalité sérielle s'exprime par un processus itératif appelé veille opérationnelle (voir figure 1) où se succèdent le recueil des données, l'intégration des informations, leur analyse, et la diffusion du renseignement (Ribaux, Genessay et Margot, 2011).

De manière surprenante, les traces matérielles sont encore peu exploitées dans ce genre de processus, malgré leur potentiel évident pour trouver des relations entre des affaires (Ribaux et al., 2003). Cette lacune s'explique notamment par la collecte et le traitement des traces orientés traditionnellement vers le tribunal, négligeant ainsi les autres aspects de l'action de sécurité.

Quoi qu'il en soit, des approches de veille opérationnelle fondées sur les informations issues de l'environnement immédiat du crime, des modes opératoires utilisés, et des traces matérielles se concrétisent par des systèmes opérationnels de suivi et d'analyse des vols en série. Ces derniers sont en cours d'évaluation mais montrent des résultats très prometteurs (Ribaux, Walsh et Margot, 2006; Rossy, Ioset, Dessimoz et Ribaux, sous presse).

La définition de l'architecture de ces processus est un enjeu capital. Quelles données sur le crime enregistrer systématiquement? Sous quelle forme?

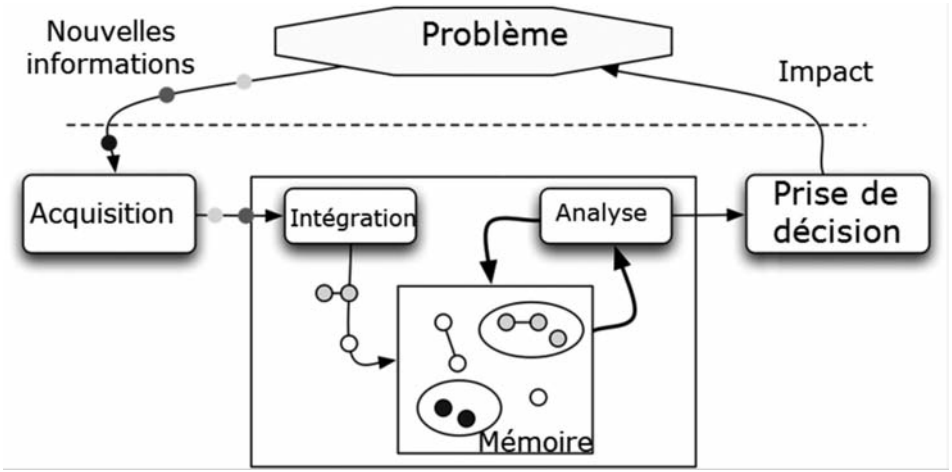


Figure 1: Processus de veille opérationnelle

Quand? Comment? Par qui? Que cherche-t-on dans les données? Quelles sont les théories qui peuvent servir à guider leur analyse? Ces questionnements requièrent d'appréhender la veille opérationnelle à l'aune de l'action de sécurité.

Patterns et modèle de police guidé par le renseignement

Ces processus de veille opérationnelle s'inscrivent dans une approche de l'action de sécurité, qui souhaite évoluer vers davantage de proactivité. Le modèle policier le plus abouti qui exprime cette volonté est celui de la police guidée par le renseignement (en anglais: intelligence-led policing) dont les facettes essentielles ont été synthétisées par Ratcliffe (2008). Dans ce modèle, l'*Interprétation* de l'environnement criminel doit *Influencer* les choix des modes d'action susceptibles de produire un *Impact* positif sur cet environnement (appelé le modèle des '3I').

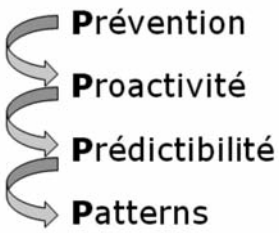


Figure 2: Modèle des 4P de Ratcliffe (2011)

L'objectif idéal est de *Prévenir* le crime en interprétant et exploitant les informations pertinentes. Toutefois, cela n'est possible qu'en développant une attitude *Proactive* qui ne peut être mise en œuvre que si le phénomène criminel présente une certaine *Prédictibilité*. Cette dernière repose entièrement sur l'existence de schémas, ou *Patterns* détectables, dont on peut supposer une réplication ultérieure. Ce modèle des 4P proposé par Ratcliffe (2011) est illustré à la figure 2. Mais ces patterns existent-ils? Si oui, ont-ils un sens? Quelle est leur nature? Peut-on les détecter, par exemple à l'aide de la veille opérationnelle? Et s'ils sont détectables, sont-ils utilisables pour élaborer et planifier une réponse? Si toutes ces questions trouvent une réponse affirmative, alors l'intérêt de l'application de techniques de data mining prend tout son sens.

Patterns criminels et data mining

L'analyse criminelle au travers de la veille opérationnelle et du data mining, veut donc détecter des patterns particuliers imprimés dans les données accessibles et engendrés par une multitude d'événements criminels. Ces patterns reflétant des schémas susceptibles de se reproduire, leur identification et leur analyse indiquent les mesures proactives les plus appropriées qui empêcheront ou perturberont leur répétition.

Plusieurs études suggèrent l'intérêt de cette alliance entre l'analyse criminelle et le data mining. En s'appuyant sur les succès obtenus dans le domaine du marketing, le data mining s'est tout d'abord appliqué à l'analyse de la criminalité financière, notamment pour les fraudes à la carte de crédit (Whitrow et al., 2009; Filipov, Mukhanov et Shchukin, 2008). Mais il s'étend également à d'autres types de criminalité, comme le crime organisé et le cybercrime (Chen *et al.*, 2003; Nissan, 2012), ou encore le trafic de stupéfiants avec la découverte de patterns dans les données chimiques des produits stupéfiants saisis (Terrettaz-Zufferey *et al.* 2006; Ratle *et al.* 2007).

Toutefois, ces études dans des champs d'applications spécifiques n'indiquent pas de schémas ou de réflexions susceptibles de délimiter la portée du data mining en analyse criminelle et d'aider à définir l'équilibre nécessaire entre l'injection de connaissances forensiques et criminologiques *a priori*, et la liberté des techniques non supervisée de data mining.

Certains patterns criminels sont faciles à détecter. Par exemple, en établissant des seuils de détection et en comparant des taux de crime survenant dans une zone cible avec les taux des précédentes semaines, des tendances deviennent détectables (Bruce, 2008). Dans d'autres situations, des modèles qui prédisent les répliques sismiques sont adaptés, par analogie, pour suivre la distribution spatio-temporelle des crimes (Mohler et al., 2011).

L'investigation du crime présente des particularités qui aident à comprendre la nature des patterns recherchés et les limites du data mining. Une activité criminelle va nécessairement engendrer des effets, qui peuvent se traduire sous forme de traces matérielles, mais aussi par extension, de caractéristiques spatio-temporelles, ou de tout autre élément résultant de l'activité. La logique indiciaire de l'investigation procède en recueillant les données issues des traces, à les analyser, puis à tenter de reconstruire ce qui s'est passé. Ce processus est imparfait, puisqu'il repose sur des données incomplètes (tous les événements ne sont pas reportés), un recueil qui ne peut pas garantir l'exhaustivité des données pertinentes (si des traces existent, il n'y a aucune garantie qu'elles soient perçues et recueillies par les investigateurs). De plus, la reconstruction procède des effets vers les causes, par abduction, dans des raisonnements qui sont approximatifs, incertains et révisables (plusieurs hypothèses peuvent expliquer les effets observés).

Le mécanisme d'analyse des données est donc plus tortueux qu'à première vue: si des patterns existent, et qu'ils sont détectés dans les données analysées, que pouvons-nous inférer par analogie sur les schémas répétitifs d'activités criminelles? Que reflètent réellement les patterns détectés? Les théories

provenant de la criminologie environnementale contribuent à répondre à ces questions.

Criminologie environnementale

Parmi les théories devenues classiques relevant des théories des opportunités ou des approches situationnelles, les activités routinières, le choix rationnel et les patterns criminels sont particulièrement importantes pour orienter l'engagement des méthodes de data mining. Ces théories indiquent combien le crime est dépendant des circonstances immédiates qui l'entourent: il ne se répartit pas aléatoirement dans l'environnement physique et spatio-temporel, et dépend de la nature et du rythme des activités sociales. On peut prétendre qu'il suit des schémas ou 'patterns' très spécifiques.

Cet ensemble de théories s'articulent autour des activités routinières de Cohen et Felson (1979). Pour qu'un crime ait lieu, il faut une rencontre dans l'espace et le temps de trois éléments: un auteur motivé, une cible propice et une absence de gardien capable. La variété des situations criminogènes très spécifiques caractérisées par l'affinité entre ces composants, ou 'opportunités', distinguent des schémas susceptibles de s'imprimer dans les données comme des patterns du crime.

Contrairement aux activités routinières de Cohen et Felson où la motivation de l'auteur à commettre un crime est supposée constante, la théorie du choix rationnel (Felson & Clarke, 1998) défend quant à elle le postulat que l'auteur motivé est un être rationnel capable de s'adapter aux situations. Sa décision de commettre un crime dépend de l'environnement immédiat et de son analyse coût-bénéfice de la situation. Ainsi, un individu peut décider de ne pas agir si les risques sont trop grands ou si la récompense est trop faible. La rationalité de l'auteur influence donc également la configuration spécifique des opportunités criminelles et prend part à la caractérisation de patterns criminels.

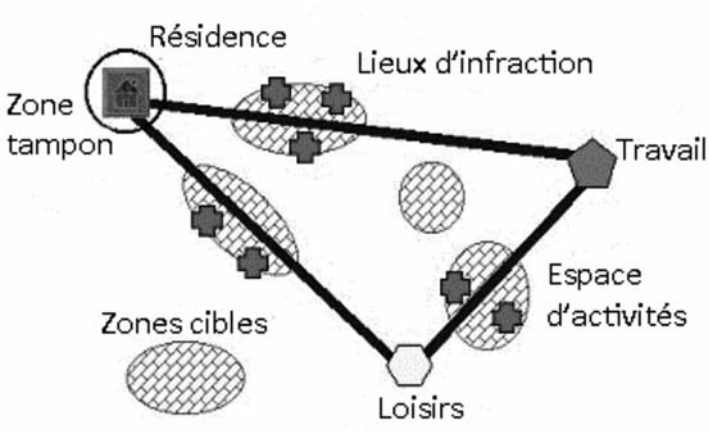


Figure 3: Trajet journalier et opportunités criminelles (traduit de Rossmo, 2000, cité par Clarke et Eck, 2005)

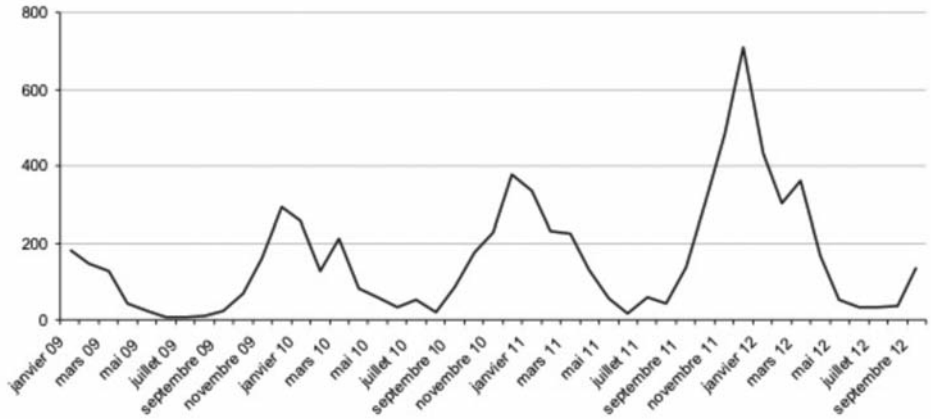
Enfin, la théorie des patterns criminels (Brantingham & Brantingham, 1990) explique comment les auteurs motivés rencontrent des cibles propices. Le trajet journalier d'un individu peut se modéliser sous la forme d'un triangle où l'individu part de chez lui pour aller au travail, puis du travail aux loisirs, et enfin des loisirs à chez lui (voir figure 3). Tout au long de ses trajets, un auteur motivé peut traverser des zones cibles qui contiennent des cibles potentielles. Ce sont surtout des espaces d'activités où les personnes peuvent vivre, travailler, s'amuser. Pour un auteur motivé, il est plus simple de commettre des crimes sur son trajet journalier, plutôt que d'effectuer un trajet spécial pour le faire.

Le modèle du triangle du crime intègre les divers éléments des théories mentionnées et appuient la compréhension et l'interprétation de ces schémas potentiels (Clarke & Eck, 2005). On y retrouve les trois pans qui représentent l'auteur, la cible/victime et le lieu, la rencontre des trois constituants l'opportunité criminelle. Chacun de ces composants peut être influencé par l'un des éléments formant le triangle extérieur: l'éducateur pour l'auteur, le gardien pour la cible/victime et le responsable pour le lieu (voir figure 4).

Ce genre de configurations inscrites dans des activités routinières répétitives, sont très spécifiques. Si le crime suit cette théorie, les données du crime reflètent ces opportunités et suivent ces schémas ou patterns. En d'autres termes, les répétitions du crime, et par la suite les patterns, sont dépendant des opportunités criminelles, elles-mêmes définies par les caractéristiques des éléments constitutifs de la situation.

Figure 4: Le triangle du crime (traduit de Clarke & Eck, 2005)

Un exemple concret sur des données policières suisses illustre le fait que le crime ne soit pas distribué aléatoirement dans l'espace et le temps, et que des patterns lus dans les informations accessibles soient susceptibles d'exprimer les spécificités des situations criminelles. Le suivi systématique de l'évolution des cambriolages d'habitations commis en soirée en Suisse romande (2) fait apparaître un pattern (des pics réguliers) particulièrement pertinent. L'activité se concentre sur les périodes hivernales et semble relever d'une fatalité statistique, bien que son intensité change au cours des années (voir graphique 1). Ce pattern spécifique peut être expliqué par une analyse situationnelle. Durant l'hiver en Suisse, la luminosité diminue très rapidement en fin de journée (selon la période, dès 17 heures), ce qui demande aux résidents d'enclencher l'éclairage pour vaquer à leurs occupations. Ainsi, le cambrioleur à la recherche d'une cible dans une zone résidentielle peut utiliser l'éclairage comme un indicateur relativement



Graphique 1: Evolution des cambriolages d'habitations en soirée en Suisse romande de 2009 à 2012 (source: PICAR (3))

fiable de présence dans une villa ou un appartement. Cette méthode de sélection de la cible vise à réduire les possibilités d'une confrontation risquée avec les occupants. Elle suit donc un schéma rationnel dont la réalisation n'est possible qu'en soirée durant l'hiver, et dans certaines zones d'habitations uniquement.

Des patterns aux connaissances

Si la criminologie environnementale nous procure une justification sur l'existence de patterns produits par une diversité d'activités criminelles répétitives, il s'agit dorénavant de les détecter et de leur donner du sens afin de les utiliser à

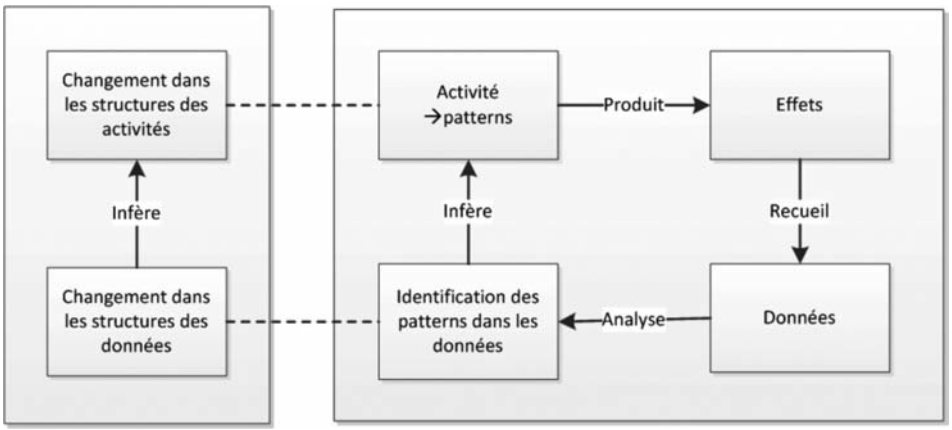


Figure 5: Hypothèse sur les changements dans les structures d'une activité criminelle

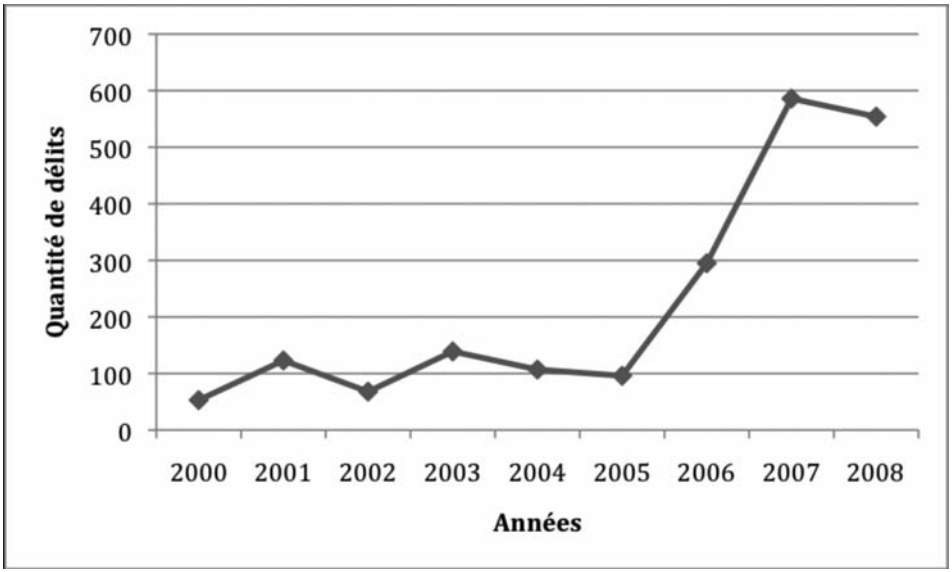
des fins de renseignement. Comme nous l'avons mentionné précédemment, cette opération suit la logique indiciaire de l'investigation qui consiste à inférer une activité criminelle en observant la structure des données qu'elle a engendrées. A partir de ce postulat, un changement dans les structures des données est susceptible de refléter une modification dans les structures des activités criminelles (voir figure 5). Encore faut-il savoir si un tel changement est décelable par l'analyse des données.

Détection des patterns

Le triangle du crime contribue à résoudre cette question: l'opportunité criminelle présente, le long de ses trois éléments constitutifs, des répétitions significatives suffisamment nombreuses pour être décelables dans les données. Selon Boba (2009), il faut penser le triangle selon le principe de Pareto, aussi appelé la règle des 80/20 (4): une minorité d'auteurs prolifiques commettent une grande majorité des crimes (Wolfgang, Figlio et Sellin, 1972; Heaton, 2000), très peu d'endroits génèrent une forte proportion des crimes (Sherman, Gartin et Buerger, 1989), tout comme un nombre restreint de victimes concentrent une grande partie des délits (Weisel, 2005). Cette approche stimule l'idée qu'une action portée sur un nombre limité d'auteurs prolifiques, de victimes répétées ou de lieux criminogènes, devrait avoir un impact significatif sur le problème global. Cette idée est confirmée par plusieurs études qui montrent l'efficacité de mesures préventives agissant sur la minorité des causes et qui auraient un impact positif sur la majorité des problèmes (Sherman et al., 1998, Chilvers and Weatherburn, 2001).

Au-delà de son orientation vers la résolution de problèmes, cette approche justifie non seulement l'existence de patterns dans les données collectées, mais donne aussi, par l'ampleur supposée de ces répétitions criminelles, de sérieux espoirs de les découvrir par une analyse ciblée des informations. Lorsqu'un nouvel auteur, ou un nouveau groupe d'auteurs, prolifique débute une activité criminelle spécifique, elle doit donc s'imprimer dans les données de manière significative. En d'autres termes, elle modifie significativement la structure de l'ensemble des activités observées, et par la suite, des informations collectées. Ces changements sont ensuite interprétés dans la logique de l'investigation qui consiste à remonter des effets (le changement de structure de données) vers les causes (comment l'expliquer). Sachant que plusieurs causes peuvent expliquer les mêmes effets, la nature 'anormale', ou inexpliquée par d'autres causes triviales, de ces changements de structure retiendront particulièrement l'attention.

Plus ces auteurs agissent de manière homogène dans l'exécution d'un délit, plus les changements deviendront perceptibles et interprétables à condition que les caractéristiques pertinentes soient relevées dans les données collectées. L'activité délictueuse des ressortissants géorgiens dans le canton de Vaud en Suisse (Azzola, 2010) constitue une illustration particulière de cette possibilité. Une augmentation significative des cambriolages diurnes d'appartement par l'arrachage du cylindre de la porte d'entrée était imprimée dans les données



Graphique 2: Evolution des cambriolages diurnes d'appartements par arrachage de cylindre (N= 2021) dans le canton de Vaud en Suisse (Azzola, 2010)

depuis 2006 (voir graphique 2). Cette anomalie fut détectée et expliquée par l'apparition de nouveaux groupes de cambrioleurs qui se sont installés dans cette niche particulière de cambriolages.

L'apparition d'auteurs prolifiques opérant dans des situations spécifiques a eu un poids important dans la répartition statistique de ce type précis de cambriolages d'habitation et a donc provoqué un changement significatif dans la structure des données.

On comprend dès lors mieux les possibilités de détection de patterns particuliers dans des tendances criminelles exprimant par exemple l'arrivée de nouveaux auteurs prolifiques. Cette détection peut résulter d'un modèle de banque de données adapté qui accueille et structure les données, ainsi que par l'exploitation d'outils d'exploration et de visualisation des informations orientée par la nature des patterns recherchés. Par exemple, les points chauds de la criminalité sont produits au moyen de systèmes d'information géographiques (SIG), et les tendances sont découvertes par des représentations graphiques et des visualisations chronologiques.

Apport du data mining

Ces instruments bien définis et exploités qui s'inscrivent dans des schémas de raisonnement suffisamment explicités sont susceptibles d'apporter des progrès considérables dans la détection des patterns pertinents (Ribaux et Margot, 1999). Ils nécessitent toutefois une intervention humaine très offensive, de l'encodage des informations jusqu'à leur analyse. Dans la réalité des services de

renseignement criminel, l'encodage des informations et la préparation des produits du renseignement ont tendance à consommer toutes les ressources. Il en reste peu pour l'exploration et l'analyse des informations, ainsi que pour traquer des patterns dans les temps exigés à la fois par l'évolution des phénomènes de criminalité et par les décideurs qui requièrent les produits à un rythme imposé par l'organisation. De plus, l'analyse est entièrement orientée par l'opérateur, ses connaissances, sa motivation et sa compréhension de la situation criminelle courante. Ces contraintes et ces biais rendent la détection des patterns souvent trop lente et l'encodage des informations montre parfois des disparités entre les analystes. En conséquence, le pattern est susceptible de rester invisible lorsque les informations sont mal structurées et ce n'est qu'une fois la tendance, ou le regroupement géographique, évident qu'il est détecté et pris en compte dans les différentes phases de la résolution des problèmes.

Les avantages potentiels pour rendre plus actives et 'intelligentes' les banques de données par le data mining présentent des formes diverses:

- améliorer la rapidité, la validité et la fiabilité de l'encodage;
- détecter plus rapidement des patterns pertinents;
- procurer des résultats pertinents inattendus pour l'analyste par une exploration à différents niveaux de détail et selon toutes les dimensions enregistrées.
- permettre une gestion multitâche à même de réduire le temps d'analyse des informations.

Ces objectifs demandent de laisser une relative liberté dans l'application des algorithmes de data mining et de ne pas trop les diriger afin d'éviter l'effet tunnel. Cet effet, qui fait référence à une perte de la vision périphérique, peut se traduire dans notre contexte par une vision circonscrite et limitée de l'objet d'étude. Dès lors, pour contrôler ces effets, les techniques relevant de l'apprentissage non-supervisé en data mining peuvent se montrer utiles, étant donné que ce type de technique ne prend pas en compte des résultats préétablis. Mais à l'inverse, une application qui n'oriente pas la recherche vers des zones prometteuses risque de ne produire que des patterns évidents, déjà bien identifiés ou qui n'ont pas de sens en termes de renseignement. Leur production risque de noyer les opérateurs de schémas et alertes inutiles, produisant ainsi l'effet inverse de celui qui est recherché. Nous revenons donc encore une fois sur cette équilibre précaire entre application supervisée et non-supervisée des techniques de data mining.

Une fois les patterns identifiés, détectés, et interprétés, il convient de les utiliser de manière adéquate en vue de produire du renseignement. On peut dès lors parler d'action préventive au sens du modèle des 4P de Ratcliffe (2008). Car la connaissance extraite à partir des patterns observés permet non seulement d'optimiser les investigations criminelles, mais surtout à orienter des stratégies proactives. En effet, en arrivant à détecter des tendances suffisamment tôt et à interpréter les patterns qui les causent, il est souvent possible d'élaborer des mesures préventives, notamment en orientant la gestion des patrouilles, en informant la population concernée, tout en s'inscrivant dans des stratégies à

plus long terme afin de résoudre les problèmes identifiés. Par exemple, les approches situationnelles en criminologie ont permis le développement des techniques de prévention situationnelle (Clarke, 1995) qui agissent sur les lieux et les cibles potentiels afin de réduire les opportunités criminelles plutôt que de chercher à détecter les auteurs. Nous nous retrouvons dans le cycle itératif du processus de veille opérationnelle, puisque c'est la prise de décision concernant des mesures préventives qui va impacter les problèmes identifiés. Ceux-ci sont non seulement traités de manière réactive, mais surtout de manière préventive en utilisant un arsenal de mesure allant au-delà de l'unique action policière.

Etude particulière de l'apport du data mining dans le suivi systématique des cambriolages d'habitations

Les questions générales posées précédemment peuvent être spécifiées dans le suivi systématique des cambriolages. En effet, malgré la présence de tendances saisonnières qui peuvent être aisément suivies une fois les patterns identifiés, la criminalité reste en constante évolution. Une mise à jour des connaissances est nécessaire afin de produire du renseignement utile et adapté. Il ne s'agit pas uniquement de se contenter de la mise en place d'outils tels que des banques de données ou l'utilisation de techniques de data mining. Ces outils expriment leur plein potentiel lorsqu'ils sont accompagnés d'une méthodologie appropriée afin d'intégrer le processus itératif de la veille opérationnelle.

Exemple en Suisse: le CICOP et PICAR

La position stratégique du renseignement criminel en Suisse romande procure une plateforme adéquate pour étudier sur un processus concret le potentiel et limites du data mining. La Suisse est un pays fédéraliste qui a 26 corps de police cantonale. Les centres régionaux d'analyse de la criminalité, comme le CICOP (*Concept Intercantonal de Coordination Opérationnelle et Préventive*) pour la Suisse romande, sont chargés de détecter et suivre la délinquance sérielle au-delà des frontières cantonales. Cette structure utilise la banque de données PICAR (*Plateforme d'Information du CICOP pour l'Analyse et le Renseignement*) qui recense tous les évènements rapportés et d'intérêt général (5) sur ce territoire. Loin d'être une simple base de données passive, PICAR s'accompagne d'une méthodologie d'analyse qui relève de la veille opérationnelle décrite précédemment (Ribaux, Genessay et Margot, 2011). Les événements enregistrés dans PICAR sont traités quotidiennement par les analystes du CICOP qui, à la suite de cela, produisent du renseignement utile et rapide. Leurs raisonnements sont basés sur des structures d'inférence comme par exemple la révision de séries criminelles (voir figure 6). Une des particularités de cette méthodologie est qu'elle s'appuie sur l'utilisation de traces matérielles pour détecter des liens entre les cas (Rossy et al., sous presse; Ribaux et Margot 1999).

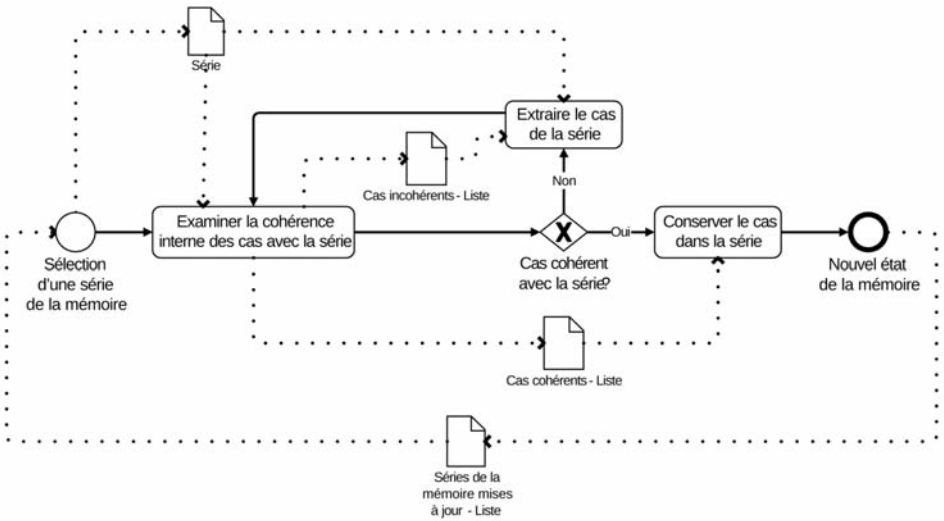


Figure 6: Structure d'inférence pour la révision de série à l'aide d'un processus BPMN (6), adapté de Ribaux et Margot (1999).

Parmi les types d'événements figurant dans PICAR, les cambriolages d'habitations sont particulièrement pris en compte par les analystes du CICOP (Ribaux et Birrer, 2008; Birrer, 2010). Hormis leur caractère éminemment sériel ou itinérant, ces cambriolages représentent surtout une grande part du volume de la criminalité en Suisse, et sont relativement bien appréhendés par les analystes du CICOP. Ces derniers utilisent des codes phénomènes afin de définir une situation criminelle présentant des patterns spécifiques. Ces codes CICOP sont directement liés aux approches situationnelles en criminologie (Birrer, 2010) et permettent une analyse plus adapté de la criminalité que la catégorisation légale traditionnelle. Cette classification se base entre autre sur le mode opératoire, sur la cible (victime ou butin), et sur le moment de la journée. Par exemple, le code CICOP «*GIORNO CILINDRO*» fait référence à un cambriolage commis durant la journée en arrachant le cylindre de la porte palière. Cette classification va donc au-delà de l'idée de catégoriser un acte criminel comme le cambriolage selon le mode opératoire (Vollmer, 1919; Birkett, 1989), et cherche à surmonter les limites exprimées par Ratcliffe (2008) en s'inspirant de la spécificité des situations criminelles (Birrer, 2010).

Recherche actuelle

Dans l'optique de cerner la contribution potentielle des techniques de data mining dans le cadre du renseignement criminel, une recherche est actuellement menée en Suisse conjointement par l'Institut de Police Scientifique de l'Université de Lausanne et l'Institut du Management de l'Information de l'Université de Neuchâtel. Il s'agit d'une recherche interdisciplinaire combinant des méthodes forensiques, criminologiques et computationnelles, et visant à

développer un cadre dans lequel les méthodes de data mining, orientées par les processus de renseignement criminel et forensiques, prennent une part active à la découverte et l'interprétation de patterns imprimés dans les données, allant jusqu'à développer de nouvelles connaissances. Réalisé en collaboration avec la Police cantonale vaudoise, la première phase de cette recherche consiste à se concentrer sur les données de cambriolages d'habitations dans le canton de Vaud et a pour principaux objectifs, la classification des cambriolages, la découverte de nouveaux phénomènes et la détection de tendances temporelles.

La première étape consiste à formaliser les processus identifiés en renseignement criminel à l'aide d'une notation standardisée appelée Business Process Model and Notation (Object Management Group, n.d.). Ensuite, différentes méthodes de data mining sont testées sur les données, puis évaluées en les confrontant avec les phénomènes déjà connus par les services d'analyse de la police. Enfin, une analyse criminologique vérifie la cohérence des résultats obtenus avec les principales théories situationnelles en criminologie.

Les premiers résultats concernent la classification automatique des cambriolages d'habitations en codes prédéfinis (code CICOP), à partir des données de base. L'avantage de cette (semi-) automatisation serait d'aider les analystes à effectuer cette opération sur les quelques dizaines de nouveaux cas quotidiens à codifier, en leur procurant ainsi un gain de temps appréciable.

Le tableau 1 montre les résultats de l'application d'une classification automatique à l'aide d'un algorithme classique de réseau neuronal appelé 'perceptron multicouches'. L'échantillon est constitué de tous les cambriolages d'habitations enregistrés dans PICAR en 2008 sur le canton de Vaud en Suisse (N= 4171). Les variables indépendantes prises en compte sont *le type de lieu, la voie d'introduction, le mode opératoire, le jour de la semaine, la saison, le moment de la journée,* et enfin *si le jour est férié ou non*. Ces variables ont pu être identifiées en étudiant les processus formalisés de raisonnement utilisés par les analystes. Cette examen indique que la variable cible, le code CICOP, peut être déterminée par les valeurs de ces variables.

L'algorithme demande une procédure d'entraînement et de test. Il apprend à classifier sur les 2/3 des données (n= 2894), puis il est testé sur les données restantes (n= 1277). Le tableau 1 nous indique le taux de précision de l'algorithme par rapport à la classification manuelle des événements. Il est remarquable de constater que ce taux est relativement élevé pour la majorité des catégories. Les classifications les moins précises («Giorno» et «Notte») sont des catégories générales, caractérisées uniquement par la période de la journée durant laquelle s'est déroulé le cas, ce qui peut expliquer la faible performance de l'algorithme. Cette expérimentation initiale reste néanmoins exploratoire en regard des objectifs plus généraux du projet sur le potentiel du data mining dans de tels processus.

Elle indique néanmoins la possibilité de soutenir par un moyen automatisé l'encodage des données, et de procurer ainsi un gain de temps non négligeable dans la mise en œuvre du processus. Cette expérimentation ouvre aussi des

Code CICOP	*Réseau neuronal (7)*
GIORNO *(cambriolage de jour sans détail)*	0.0%
GIORNO CHIAVE *(cambriolage de jour avec introduction clandestine avec clé trouvée)*	90.9%
GIORNO CILINDRO *(cambriolage de jour par la porte avec arrachage du cylindre)*	98.9%
GIORNO EPAULEE *(cambriolage de jour par la porte à coups d'épaule, de pieds, ou de vive force)*	92.0%
GIORNO FINESTRA *(cambriolage de jour par la fenêtre)*	88.9%
GIORNO PIATTO *(cambriolage de jour par la porte avec outil plat)*	98.2%
HALL *(cambriolage de jour par introduction clandestine et fouille du hall d'entrée)*	100.0%
NOTTE *(cambriolage de nuit sans détail)*	18.9%
NOTTE CHIGNOLE *(cambriolage de nuit par la fenêtre avec une chignole)*	100.0%
NOTTE CILINDRO *(cambriolage de nuit par la porte avec arrachage du cylindre)*	57.1%
NOTTE FINESTRA *(cambriolage de nuit par la fenêtre)*	67.6%
SERA *(cambriolage du soir sans détail)*	65.7%
SERA BLOKO *(cambriolage du soir en bloquant la porte d'entrée)*	60.0%
Non classifié	82.6%
Total	*84.5%*

Tableau 1: Taux de précision d'un réseau neuronal dans la classification de cambriolages d'habitations (n= 1277)

perspectives dans l'utilisation d'approches qui ne nécessitent pas une classification déterministe des cas, mais qui mémorise plutôt des degrés d'appartenance à une ou plusieurs catégories. Cette approche offre évidemment davantage de souplesse pour l'analyse, car elle laisse ouverte la possibilité d'interpréter les cas sous différentes hypothèses. Ce genre de modèles seront développés dans la prochaine phase du projet.

Conclusion

Les outils de data mining paraissent appropriés pour soutenir le processus de renseignement criminel et la veille opérationnelle. Mais, au-delà de nos premières expérimentations liées à la classification, là où cette contribution sera la plus attendue est bien évidemment dans la détection de patterns promise par les théories situationnelles, tels que des tendances ou des schémas extraordi-

naires susceptibles d'indiquer l'activité d'un même auteur ou groupe d'auteurs, des points chauds ou des traces de victimisations répétées. La détection et la compréhension par la logique de l'investigation de phénomènes méconnus est elle aussi une des attentes principales.

Notre démarche consiste donc à prendre en considération les contraintes et spécificités liées à la production de renseignement criminel dans l'application du data mining. Ces techniques computationnelles ne peuvent pas être réduites à une baguette magique qu'il suffit de brandir pour extraire de la connaissance dans des ensembles de données. De la même manière qu'une banque de donnée telle que PICAR en Suisse doit être accompagnée d'une méthodologie particulière, le data mining doit être encadré en amont et en aval par les théories criminologiques et l'expérience accumulée par les praticiens afin d'orienter l'application de ces techniques.

L'équilibre à atteindre sur le degré de supervision des techniques de data mining est subtil. Ce point d'équilibre n'est ni unique, ni permanent, il dépend de l'étape du processus à laquelle s'applique le data mining et de son objectif. Un projet de recherche mené en Suisse vise à le situer et à définir son pilotage «intelligent» au travers de diverses expérimentations sur les processus impliqués dans le renseignement criminel tel que conduit au sein du CICOP. Malgré des premiers résultats prometteurs, le terrain de recherche est encore vaste, et nombreuses sont les tâches à effectuer.

Le data mining est donc potentiellement capable d'apporter de nombreux avantages en termes de célérité et de complétude des analyses, et permettrait ainsi de se libérer des *a priori* pouvant restreindre la portée de la veille opérationnelle. Cependant, son application doit être encadrée par une méthodologie solide dont les contours restent encore à définir afin d'éviter des dérives utopistes sur ses promesses.

Remerciements

Les auteurs voudraient remercier la police de sûreté et son Chef Alexandre Girod de la Police cantonale vaudoise, ainsi que les analystes du CICOP Sylvain Ioset et Damien Dessimoz pour leur coopération et les données mises à disposition.

Références

Azzola, A. (2010). *Délinquance des ressortissants géorgiens dans le canton de Vaud à partir de données policières entre 2000 et 2008.* (mémoire de maîtrise non publié). Université de Lausanne, Suisse.

Birkett, J. (1989). Scientific scene linking. *Journal of the Forensic Science Society*, *29*(4), 271-284.

Birrer, S. (2010). *Analyse systématique et permanente de la délinquance sérielle: Place des statistiques criminelles; apport des approches situationnelles pour un système de classification; perspectives en matière de coopération*. (thèse de doctorat, Université de Lausanne, Suisse). Récupéré du site de l'Ecole des Sciences Criminelles de l'Université de Lausanne: http://www.unil.ch/esc/page18345.html

Boba, R. (2009). *Crime Analysis with Crime Mapping.* California, USA: Sage.

Brantingham, P. J. et Brantingham, P. L. (1990). Situational crime prevention in practice. *Canadian Journal of Criminology*, *32*, 17-40.

Cao, L. (2008). Domain Driven Data Mining (D^3M). Dans F. Bonchi et al. (eds.), *IEEE International Conference on Data Mining Workshops* (*ICDM Workshops 2008)* (p. 74-76). Récupéré du site: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4733896

Cao, L., Yu, P. S., Zhang, C. et Zhao, Y. (2010).*Domain Driven Data Mining*. New York, USA: Springer.

Bruce, C. W. (2008). The Patrol Route Monitor: a Modern Approach to Threshold Analysis. *International Association of Crime Analysts.*

Chen, H., Chung, W., Qin, Y., Chau, M., Jie Xu, J., Wang, G., Zheng, R. et Atabakhsh, H. (2003). Crime Data Mining: An Overview and Case Studies. Dans *proceedings National Conference on Digital Government Research 2003,* University of Arizona.

Chilvers, M. et Weatherburn, D. (2001). Operation and crime review panels: Their impact on break and enter. *Crime and Justice Statistics Bureau Brief, Avril 2001,* Sydney: NSW Bureau of Crime Statistics and Research.

Clarke, R. V. et Eck, J. E. (2005). *Crime Analysis for Problem Solvers In 60 Small Steps*. USA: Center for Problem Oriented Policing.

Clarke, R. V. (1995). Les technologies de la prévention situationnelle. *Les Cahiers de la sécurité intérieure, 21*(3), 101-112.

Cohen, L. E. et Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588-608.

Fayyad, U., Piatetsky-Shapiro, G. et Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*, 37-54.

Felson, M. et Clarke, R. V. (1998). *Opportunity makes the thief: Practical theory for crime prevention*. Police Pesearch Series Paper. Londres: Home Office.

Filipov, V., Mukhanov, L. et Shchukin, B. (2008). Transaction aggregation as strategy for credit card fraud detection. *Cybernetic Intelligent Systems CIS 2008 .* Présenté à 7th IEEE International Conference on Cybernetic intelligent Systems, Londres, UK.

Heaton, R. (2000). The prospects for intelligence led policing: Some Historical and quantitative considerations. *Policing and Society, 9*(4), 337-355.

Kahneman, D. (2011). *Thinking Fast and Slow*. Londres: Allen Lane.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. et Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association, 106*(493), 100-108.

Nissan, E. (2012). An Overview of Data Mining for Combating Crime. *Applied Artificial Intelligence: An International Journal, 26*(8), 760-786.

Object Management Group (n.d.). *Business Process Model and Notation*. Récupéré du site: http://www.bpmn.org/

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

Ratcliffe, J.H. (2011). Intelligence-led policing: Anticipating risk and influencing action. in Wright, R, Morehouse, B, Peterson, MB & Palmieri, L (Eds). *Criminal Intelligence for the 21st Century*, IALEIA, 206-220.

Ratle, F., C. Gagné, A.-L. Terrettaz-Zufferey, M. Khanevski, P. Esseiva et O. Ribaux (2007). Advanced Clustering Methods for Mining Chemical Databases in Forensic Science. *Chemometrics and Intelligent laboratory Systems*, *90*, 122-131.

Ribaux, O. (1997). *La recherche et la gestion des liens dans l'investigation criminelle: le cas particulier du cambriolage*. (thèse de doctorat non publiée). Université de Lausanne, Suisse.

Ribaux, O. et Birrer, S. (2008). Système de suivi et d'analyse des cambriolages appliqué dans des polices suisses. *Erstes Zürcher Präventionsforum: Kommunale Kriminalprävention Crime Mapping Einbruchskriminalität*, Europa Institut Zürich, 189-205.

Ribaux, O., Genessay, T. et Margot, P. (2011). Les processus de veille opérationnelle et science forensique. In S. Leman-Langlois (Éd.), *Sphères de surveillance* (Les Presses de l'Université de Montréal., p. 137-158). Montréal.

Ribaux, O., Girod, A., Walsh, S. J., Margot, P., Mizrahi, S. et Clivaz, V. (2003). Forensic intelligence and crime analysis. *Law, Probability and Risk*, *2*, 47-60.

Ribaux, O. et Margot, P. (1999). Inference structures for crime analysis and intelligence: the example of burglary using forensic science data. *Forensic Science International*, *100*(3), 193-210.

Ribaux, O., Taroni, F. et Margot, P. (1995). La recherche et la gestion des liens dans l'investigation criminelle: une étape vers l'exploitation systématique des données de police. *Revue internationale de Criminologie et de Police Technique*, (2), 229-242.

Ribaux, O., Walsh, S. et Margot, P. (2006). The contribution of forensic science to crime analysis and investigation: Forensic intelligence. *Forensic Science International*, *156*, 171-181.

Rossmo, K. (2000). *Geographic Profiling.* Boca Raton, FL: CRC Press (d'après Clarke et Eck, 2005).

Rossy, Q., Ioset, S., Dessimoz, D. et Ribaux, O. (sous presse). Integrating forensic information in a crime intelligence database. *Forensic Science International*.

Sherman, L. W., Gartin, P. R. et Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*(1), 27-56.

Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. et Bushway, S. (1998). *Preventing Crime: What works, what doesn't, what's promising*. Washington DC: National Institute of Justice.

Terrettaz-Zufferey, A.-L., F. Ratle, O. Ribaux, P. Esseiva et M. Kanevski (2006). Assessment of Data Mining Methods for Forensic Case Data Analysis. *Journal of Criminal Justice and Security (Varstovoslovje), Special issue no 3-4*, 350-355.

Tufféry, S. (2007). *Data Mining et statistique décisionnelle: L'intelligence des données.* Paris: TECHNIP.

Vollmer, A. (1919). Revision of the Atcherley Modus Operandi System. *Journal of the American Institute of Criminal Law and Criminology*, *10*(2), 229.

Weisel, D. L. (2005). Analyzing Repeat Victimization. *Problem-Oriented Guides for Police: Problem-Solving Tools Series, 4*. USA: Center for Problem Oriented Policing.

Whitrow, C., Hand, D. J., Juszczak, P., Weston, D. et Adams, N. M. (2009). Transaction aggregation as strategy for credit card fraud detection. *Journal of Data Mining and Knowledge Discovery*, *18*(1), 30-55.

Wolfgang, M. E., Figlio, R. M. et Sellin, T. (1972). *Delinquency in a Birth Cohort*. Chicago, IL: University of Chicago Press.

Notes

1 Cette recherche est soutenue par le Fonds National Suisse de la Recherche Scientifique FNSRS No 135236

2 Partie francophone du territoire suisse

3 Plateforme d'Information du CICOP pour l'Analyse et le Renseignement

4 Il s'agit d'une proportion symbolique, selon Ratcliffe (2008), il s'agit plutôt de 60% des délits commis par 6% des auteurs.

5 Principalement les événements susceptibles de présenter un caractère sériel ou itinérant.

6 Business Process Model and Notation: notation standardisée pour représenter des processus

7 Calculé avec le logiciel SPSS