Étude critique d'un système d'analyse prédictive appliqué à la criminalité : Predpol®

Ismael Benslimane
Ismael.Benslimane@e.ujf-grenoble.fr
Sous la direction de Guillemette Reviron*
Cortecs / Université Joseph Fourrier - Grenoble

Le 18 juin 2014

Résumé. — *Nous avons réalisé une analyse critique du système commercial de prédiction policière dénommé Predpol® visant à évaluer objectivement l'efficacité des prédictions réalisées. Après avoir effectué un travail de recoupement d'informations, nous avons recueilli les données nécessaires pour répondre à notre problématique. Nos simulations numériques ont pu mettre en évidence que, sur l'étude concernant la prédiction des homicides et délits avec arme à feu à Chicago, les résultats parus dans l'article (Mohler, 2014) ne sont pas probants. En effet, avec un algorithme basique de "prédiction par meilleur rang avec points chauds dynamiques" nous obtenons des résultats équivalents. Une valeur du pAUC (Aire partielle sous la courbe) standardisée donne 0.559±0.01 pour notre étude versus 0.557±0.01 (Mohler, 2014, Fig. 1, numérisée par OPR) pour la leur, mettant in fine en question la plus-value prétendue par G. Mohler et Predpol Inc.*

I. Introduction

Cette étude a été effectuée au sein du Cortecs (COllectif de Recherche Transdisciplinaire Esprit Critique & Sciences), structure universitaire développant l'analyse critique en science. L'étude fut supervisée par Guillemette Reviron, docteure en Mathématiques, chargée de cours d'enseignements critiques à l'Université Montpellier 2.

Predpol, application commerciale de prédiction des délits en zone urbaine. — Predpol, par le biais d'une interface cartographique, affiche quotidiennement un ensemble de zones à risques (par exemple : 20 cases de 150m x 150m) pour lesquelles la probabilité qu'un délit se manifeste est importante.

Précision et évolution de la criminalité. — Predpol affirme que son algorithme est jusqu'à deux fois plus précis que les analystes spécialisés (Predpol-Inc, 2013). L'entreprise prétend également être à l'origine d'une diminution de la criminalité. Par exemple, elle communique le chiffre de 12% de diminution de la criminalité constatée dans la division de Foothill à Los Angeles après 6 mois de mise en service et compare ce chiffre à une augmentation de 0.4% dans le reste de la ville.

Un développement conséquent et un fort impact médiatique. — Nous pouvons noter que le système a été l'objet d'une promotion très importante. Des centaines de grands médias ont évoqué le sujet et le "Time Magazine" désigna Predpol comme l'une des "50 inventions de l'année 2011". Le système est mis en place dans plus d'une vingtaine de villes aux États-Unis et une au Royaume-Uni. Plusieurs articles de recherches ont été subventionnés par la National Science Foundation (DMS-0968309) ainsi que le département de recherche de la défense Étasunienne (58344-MA).

Prédiction par des modèles de processus ponctuels. — La théorie sous-jacente aux algorithmes développés est la théorie des processus ponctuels. Cette sous-branche des statistiques et des probabilités est utilisée dans des domaines variés (sismologie, épidémiologie, écologie, économétrie...) afin de modéliser des phénomènes appartenant à la géométrie stochastique. Les principaux algorithmes utilisés sont des modèles de processus ponctuels "auto-excités" (Mohler et al., 2011) ainsi que des modèles de processus ponctuels marqués (Mohler, 2014). Ceux-ci consistent à attribuer au temps t et à chaque point (x, y) de la carte, une pondération $\lambda(x, y, t)$ calculée en additionnant deux composantes, l'une évaluant des données chroniques $\mu(x, y)$ indépendantes du temps, l'autre recueillant des données dynamiques afin de modéliser par un noyau g les corrélations entre délits proches. Pour affiner la modélisation on marque $M = \{1, 2, 3, ...\}$ la catégorie du crime ; $M = 1$ étant les homicides, $M = 2$ les vols, *etc...* On obtient pour chaque point spatio-temporel :

$$\lambda(x, y, t) = \mu(x, y) + \sum_{i, t_i < t} g(x - x_i, y - y_i, t - t_i, M_i) \quad (1)$$

La somme représente la comptabilisation de tous les délits de notre historique, g est une fonction qui dans

*Remerciements : je remercie chaleureusement Guillemette Reviron pour son suivi méticuleux, ses brillants conseils et sa grande sympathie. Je remercie toute l'équipe du Cortecs pour les agréables moments passés et les renversantes discussions. Merci à Marcela Perrone (LPNC, Grenoble) pour certains aspects techniques concernant l'accès à plusieurs revues scientifiques. Enfin, je voudrais remercier particulièrement David Marsan, sismologue à l'ISTerre (CNRS, Université de Savoie) pour m'avoir accordé du temps afin de m'aider dans mes recherches.

l'étude (Mohler, 2014) est une exponentielle décroissante en espace et en temps. L'évaluation de $\lambda(x, y, t)$ est réalisée en pondérant tous les délits passés en fonction de l'ancienneté $t - t_i$ et de l'éloignement $x - x_i, y - y_i$. Plus un délit est éloigné de la zone étudié et ancien par rapport au moment de la prédiction, moins il comptera dans le calcul du poids en (x, y). Les paramètres du noyau g sont ajustés de manière empirique par des algorithmes de maximisation.

II. Méthodes

La première partie de l'étude porta sur la recherche d'informations et le recoupement de celles-ci afin d'établir un profil complet du système développé par *Predpol Inc*.

Sources de données recueillies sur Predpol et les chiffres de la criminalité. — Nous avons contacté la société *Predpol Inc* localisée à Santa Cruz, CA, États-Unis ainsi que la police du comté du Kent située au Royaume-Uni et utilisant Predpol depuis décembre 2012. Les démarches envers elles n'ont malheureusement donné aucune suite exploitable. L'ensemble des informations recueillies proviennent d'articles de recherche, de sites internet (société *Predpol Inc*, services de police), de présentations du produit, de conférences en ligne, de bases de données publiques ainsi que de rapports confidentiels anonymisés et déclassifiés.

II.1. Étude de trois cas : Los Angeles, le comté du Kent et Chicago

Foothill 2011 : étude randomisée réalisée pendant 6 mois à Los Angeles, CA, États-Unis. — Étude la plus évoquée par les médias et seule étude réellement expérimentale mise en œuvre. Pour autant, aucun article scientifique n'est paru à son sujet. Or, les résultats promus par la société Predpol Inc et par les unités de Police étant trop peu développés, ils s'avèrent inexploitables. La gestion des données criminelles est effectuée par une société privée *The Omega Group*. Cette société rend accessible au public les données géoréférencées des délits par l'interface CrimeMapping.com™. Cependant, les données sont restreintes aux délits récents (moins de six mois). Il n'a donc pas été possible d'obtenir les données concernant les six mois de l'expérimentation. Pour autant certaines analyses ont pu être faites avec des données postérieures (2014).

Kent 2012 : étude réalisée au Royaume-Uni. — Grâce à un document déclassifié [1], nous avons pu avoir accès au compte-rendu de l'expérimentation. Encore une fois, les informations concernant le protocole sont peu précises et non exploitables. Par ailleurs, les données géoréférencées sont publiques et librement disponibles [2]. Cependant, pour des raisons de confidentialité, seuls sont publiés le mois et l'année du délit et non le jour et l'heure, données nécessaires pour notre étude.

Chicago 2014 : étude théorique rétrospective. — Deux raisons nous ont amenés à nous concentrer sur le cas de Chicago. D'une part, les données sont publiques et libres d'accès [3]. La base de données contient un peu plus de 5 500 000 délits répertoriés de 2001 jusqu'à aujourd'hui avec pour champs, le type de délit, la description, les coordonnées GPS, l'heure, la date, *etc...* D'autre part, les seules études sujettes à une analyse approfondie ont été celles concernant la ville de Chicago avec un bref rapport (Predpol-Inc, 2013) et une publication (Mohler, 2014) à paraître dans l'International Journal of Forecasting de Juillet-Septembre 2014. Cette dernière publication d'une revue scientifique à comité de lecture est la continuité du premier rapport mais développe une analyse beaucoup plus complète. Nous avons été en mesure de comparer nos simulations avec celles de l'article par OPR (Otical Plot Reading).

II.2. Trois méthodes d'évaluation : la prédictibilité, l'efficacité, l'efficience

La prédictibilité : mesure concernant la précision des prédictions. — La précision du système à prédire un délit est mesurée en donnant la fraction de délits correctement prédits quotidiennement. En faisant varier le nombre de prédictions quotidiennes, on obtient une courbe similaire à une courbe ROC (Receiver Operating Characteristic), caractérisant l'efficacité d'un test prédictif en évaluant la sensibilité *versus* la spécificité.

L'efficacité, mesure de la criminalité constatée. — La seconde évaluation concerne le taux de criminalité en analysant les fluctuations du nombre de délits quotidiens avant et après la mise en place du système d'analyse prédictive. Cette dernière évaluation, qui est sûrement la plus attractive, est cependant très délicate. En effet de nombreux biais peuvent amener à penser à tort que la mise en place d'un tel système a effectivement diminué ou augmenté la criminalité. Nous pouvons dès à présent évoquer quelques-uns de ces biais comme :

- un tri des données
- des confusions de corrélation/causalité
- des problèmes de dénombrement du nombre de délits

1. Document déclassifié accessible ici : `https://www.whatdotheyknow.com/request/181341/response/454199/attach/3/13%2010%20888%20Appendix.pdf]`
2. Données disponibles ici : `http://data.police.uk`
3. Base de données téléchargeable ici : `https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2`

L'efficience, évaluation des moyens nécessaires. — Évaluation des possibilités techniques de mise en œuvre d'un tel système (ressources humaines, matérielles, informatiques, *etc...*).

Ces deux derniers points demandent des discussions approfondies qui sortent du cadre de cette étude.

II.3. Simulations et analyses effectuées : génération de prédiction et comparaison des résultats

Le but de notre étude fut donc de comparer les résultats obtenus par Predpol en terme d'efficacité de prédiction avec des méthodes d'analyse basique pour pouvoir évaluer l'efficacité propre de Predpol.

Base de données et conditions d'expérimentation. — Nous avons téléchargé la même base de données qui est décrite dans les articles (Predpol-Inc, 2013; Mohler, 2014). Celle-ci concerne uniquement les homicides et délits avec armes à feu entre 2007 et 2012. Nous avons utilisé les mêmes critères de sélection que Predpol et nous avons obtenu une base de données avec des différences négligeables, 78 887 données contre 78 852 dénombrées dans (Mohler, 2014), c'est-à-dire une erreur de 0,04%. Pour s'assurer que les conditions de simulations étaient les mêmes, nous avons ajusté notre grille (220×290 cases de surface $S_{case} = 150m \times 150m$ avec pour coordonnées Nord-Ouest (42.0243,-87.9093)) sur celle de Predpol en nous basant sur une capture d'écran présente dans l'un des rapports étudiés (Predpol-Inc, 2013).

Algorithmes concurrents développés. — Nous avons développé plusieurs algorithmes assez basiques afin d'effectuer une comparaison de leur efficacité. Ils sont dits basiques car ils ne prennent pas en compte un certain nombre de corrélations statistiques. Voici comment nous avons procédé : chaque jour a été générée une carte avec un certain nombre de zones sélectionnées comme prédiction d'un ou plusieurs délits.

Prédiction aléatoire. — Il est toujours important de comparer une prédiction à une analyse par génération aléatoire. La proportion du nombre de délits prédits par rapport au nombre de délits constatés en tirant aléatoirement N cases s'obtient par l'équation :

$$p(N) = \frac{N \times S_{case}}{S_{totale}} \tag{2}$$

Prédiction aléatoire avec points chauds. — Cette technique est une amélioration de la précédente. Cette fois-ci, on prend en compte les spécificités du terrain. Pour chaque division $r(x, y)$, on calcule le nombre de délits qui se sont produits dans ce même lieu dans l'historique $[t_1, t_2]$. On obtient donc une carte de la répartition de l'ensemble des délits enregistrés. Une fois cette pondération effectuée, nous avons tiré aléatoirement N cases en prenant en compte la pondération, ces N cases donnent nos prédictions concernant les lieux où se produiront des infractions. Après avoir essayé des algorithmes basés sur des tris de tableau, ou d'autres de sélections génétiques (appelés aussi, sélections par le jeu de la roulette), nous avons finalement trouvé un algorithme plus rapide en traitement et moins biaisé statistiquement parlant, celui de C. K. Wong and M. C. Easton (Wong and Easton, 1980). Il utilise une structure de données en arbre binaire (btree) et permet une génération en temps $O(NlogN_{tot})$ où N est le nombre de cases sélectionnées pour la prédiction et N_{tot} le nombre total de cases de notre carte. En attribuant une valeur à chaque nœud correspondant à la pondération de chaque case, le tirage s'effectue très rapidement en parcourant l'arbre binaire.

Prédiction par meilleur rang avec points chauds. — La prédiction par meilleur rang utilise les mêmes principes que les algorithmes avec points chauds pour la pondération. Cependant, c'est au niveau de la sélection que cela diffère : au lieu d'effectuer une sélection aléatoire, les zones à plus haut risque sont choisies, chaque division de la carte est triée par taux de criminalité, et les N-ièmes premières divisions sont choisies comme étant nos prédictions.

Algorithmes de prédiction

1. **Aléatoire ignare :** prédiction par tirage au hasard de N uniques cases sans prise en compte des spécificités du terrain.
2. **Aléatoire avec points chauds chroniques :** la pondération effectuée avec l'historique des délits est calculée une seule fois avec, par exemple, les trois ans précédant le début du lancement des prédictions pour l'année. Chaque jour j au cours de l'année est générée une prédiction avec ces mêmes données non actualisées.
3. **Aléatoire avec points chauds dynamiques :** contrairement à l'algorithme chronique, on met à jour notre pondération avec les données du jour $j-1$ précédant la prédiction du jour j.
4. **Meilleur rang avec points chauds chroniques :** on utilise une pondération non mise à jour. Tous les jours de l'année on prédira donc les mêmes zones.
5. **Meilleur rang avec points chauds dynamiques :** on utilise une pondération mise à jour jusqu'au jour $j-1$.

III. Résultats et discussion

Après avoir généré nos prédictions pour les années 2010-2012 avec nos différents algorithmes, et en faisant varier de 0 à 500 (sur 26900) le nombre de cases prédites chaque jour, nous obtenons :

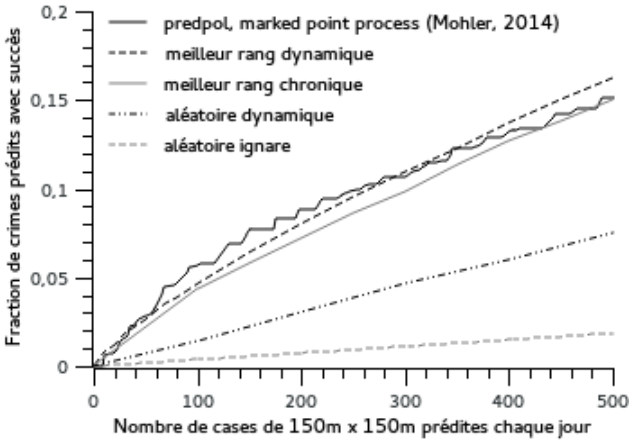

Figure 1 – *Fraction de délits prédits avec succès entre 2010 et 2012 versus nombre de cases prédites chaque jour. — Comparaison avec la publication (Mohler, 2014, Fig. 1.)*

Outre le fait que ces résultats montrent l'augmentation de la prédiction lorsque l'on prend en compte les données récentes (dynamique) par rapport au données antérieures (chronique), nous pouvons aussi constater que la sélection par meilleur rang est plus performante que l'aléatoire par point chaud. L'algorithme aléatoire ignare donne les plus mauvais résultats, la droite linéaire est définie par l'eq. 2.

Notre algorithme le plus performant (meilleur rang dynamique) obtient des scores de prédiction très proches de la courbe (Mohler, 2014, Fig.1. "marked point process"). Une bonne caractéristique de ce type de courbe ROC est le calcul du pAUC (Aire partielle sous la courbe) standardisé ; valeur comprise entre 0.5 (aléatoire) et 1 (100% juste). Dans l'intervalle [0,500], on obtient un pAUC de 0.541 pour l'algorithme de meilleur rang dynamique contre 0.542 pour le processus marqué. Sur l'intervalle [250,500] préconisée pour un déploiement à Chicago (Mohler, 2014), l'efficacité propre du processus marqué n'est de nouveau pas mise en évidence avec un pAUC de 0.557 contre 0.559 pour notre plus performant algorithme. L' incertitude estimée est de ± 0.01.

Répartition spatiale et dynamique des délits. — La courbe (Fig. 2) montrant la fraction de délits située dans une fraction d'espace nous indique le fait que la répartition est très inégale. Cette répartition permet d'expliquer en partie l'efficacité d'un algorithme de meilleur rang comparé à des techniques plus fines, et relativise les affirmations de Predpol promouvant le fait de prédire 50% des délits en pointant 10.3% de la surface de la ville (Predpol-Inc, 2013).

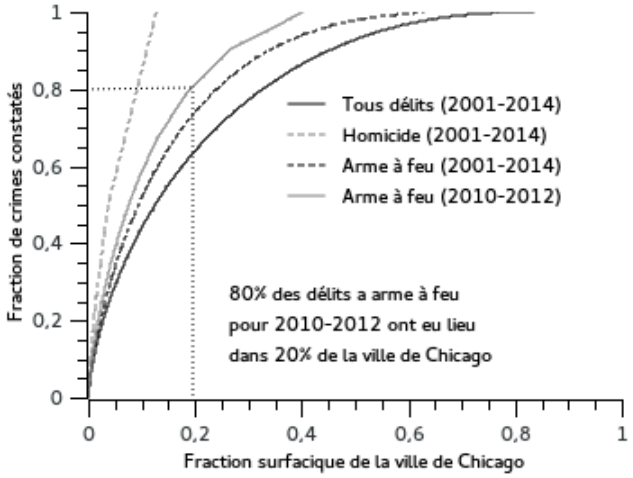

Figure 2 – *Répartition des délits — Fraction de délit versus fraction surfacique de la ville de Chicago*

IV. Conclusion

Les résultats obtenus avec l'analyse rétrospective, nous permettent d'avoir de sérieux doutes sur l'efficacité propre de Predpol en condition réelle. Ceci nous amène à continuer notre investigation. Nous allons recontacter les auteurs des études (Predpol-Inc, 2013; Mohler, 2014) afin de les confronter à nos résultats. D'autres suites seront envisageables selon leur réponse. Au-delà des aspects les plus techniques, nous pouvons nous demander, si la démarche de Predpol ne désyncrétise pas la question de la criminalité en laissant penser qu'il suffit de prédire les délits pour en diminuer le nombre. Predpol et sa médiatisation véhiculent ainsi une idée répandue, "simple" et séduisante oubliant *de facto* les facteurs sociologiques amenant aux comportements délictueux. La question fondamentale des inégalités de répartition des richesses est, par exemple, rarement débattue. En effet, cette réflexion est beaucoup plus impliquante à long terme qu'un logiciel spectaculaire.

Références

George Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3) :491–497, 2014.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

Predpol-Inc. PredPol Predicts Gun Violence. `http://www.predpol.com/wp-content/uploads/2013/06/predpol_gun-violence.pdf`, 2013.

Chak-Kuen Wong and Malcolm C. Easton. An efficient method for weighted sampling without replacement. *SIAM Journal on Computing*, 9(1) :111–113, 1980.