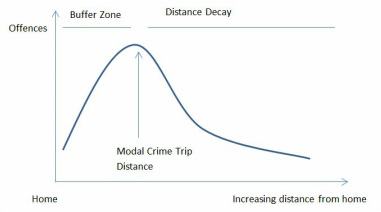
**Sujet de statistiques appliquées – orientation Data Science & Actuariat**

**2014-2015**

**Modèles espace-temps du risque de cambriolage**

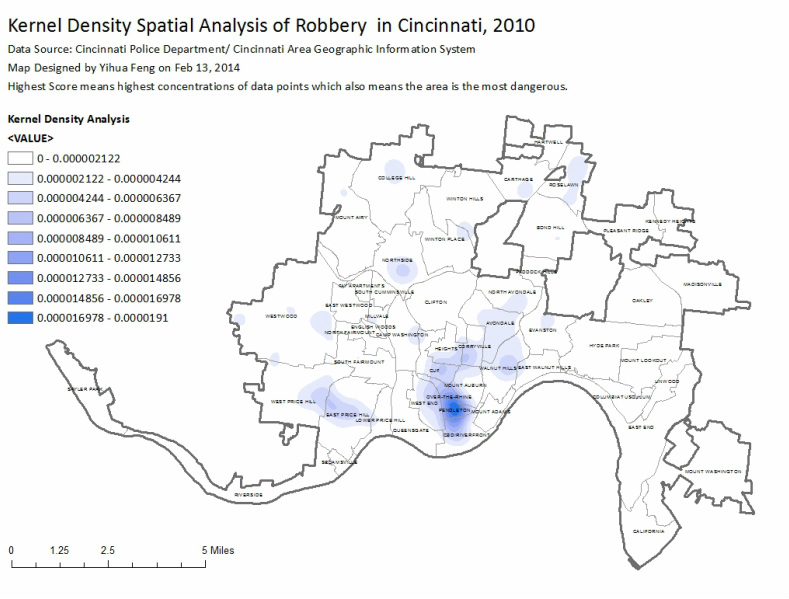
**Application à Paris et Bruxelles à partir des données d’assurance**

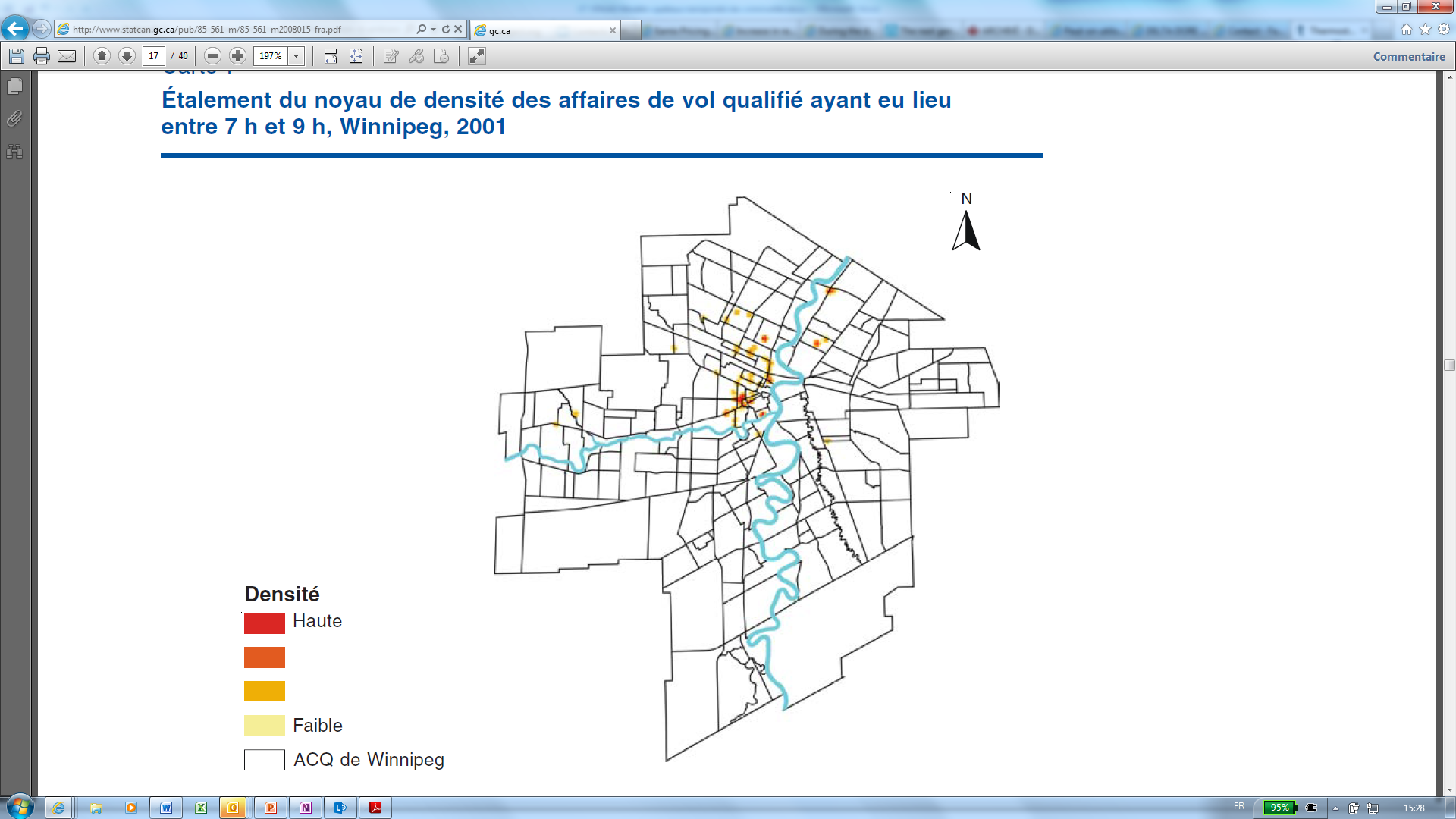
Le but de ce projet de « statistiques appliquées » est d’étudier les caractéristiques spatiales et temporelles de la criminalité vol afin de mettre en place un système d’alerte au vol et de mieux appréhender ce risque. En effet, l’influence spatiale de la criminalité est bien établie même si elle peut intervenir de façon complexe (par exemple, la sur-criminalité n’est pas directement liée à des caractéristiques socio-économiques mais à la proximité de certaines zones riches avec des zones plus défavorisées). Par exemple, selon Rengert et al 1999; Rossmo, 2000, la sinistralité est maximale à quelque distance du lieu d’habitation du voleur (Distance Decay Effect ), avec une distance médiane de 2.6 km (distance euclidienne - Montréal, 2001):



La statapp aura donc pour but d’établir un modèle spatiale basée sur la théorie de criminologie / cambriolage et sur l’application aux données assurantielles, puis d’étudier ce modèle de façon dynamique : comment en particulier détecter les phénomènes nouveaux de vol afin de mieux les prévenir ?

Pratiquement, le sujet pourra être découpé de la façon suivante :

* Revue de la littérature théorique sur la criminalité / vol (théorie des opportunités criminelles,…)
* Proposition d’un ou plusieurs modèles spatiaux à tester sur les données d’AXA tenant compte de la littérature théorique ou d’approches purement statistiques (modèle d’Auto-régression conditionnelle (CAR)…).
* Application de ces modèles sur les données AXA (sinistre et police géocodé) France et Belge
* Recherche de données externes (socio-démographiques / criminelles) permettant d’améliorer le modèle.
* 
* Intégration de données supplémentaires sur les caractéristique du sinistre (heure, type de vol…) pour affiner le modèle spatiale



* Discussion sur les différents modèles statistiques et théoriques
* Puis extension de ces modèles en intégrant l’approche dynamique. Cette partie pourra être l’occasion de tester l’efficacité des modèles alternatifs non paramétriques (gradient boosting),….
* Création d’un modèle prédictif de vol en fonction de cette approche dynamique afin de pouvoir prévenir en amont du risque de vol.

Encadrement par deux anciens élèves de l’ENSAE (un junior et un top manager promo 1990)

Littérature – quelques pistes d’introduction :

* <http://www98.griffith.edu.au/dspace/bitstream/handle/10072/24437/49610_1.pdf?sequence=1> : a cross national assessment of space-time patterns of burglary risk, Shane D Johnson et al - 1998
* Computational Actuarial Science with R, Arthur Charpentier (2014) : chapitre sur la gestion statistique des données spatiales sous R.
* <http://www.statcan.gc.ca/pub/85-561-m/85-561-m2008015-fra.htm> : analyse spatiale de la criminologie au Canada
* <http://evasdatavisualization.weebly.com/project-1.html> : Exemple sur la ville de Cincinnati example
* <http://www.math.u-psud.fr/~stafav/IMG/pdf/Statistique_Spatiale.pdf> : cours de statistique spatiale
* R : documentation des packages spatiaux geoR, gstat, RandomFields, spatial, spatstat, spdep

Pour plus de contacts : [guillaume.gorge@axa.com](mailto:guillaume.gorge@axa.com) et [phileas.condemine@axa.com](mailto:phileas.condemine@axa.com)

**Rdv avec Anne-Laure Le Gallo, Doan Nguyen et Julien Callard - AXA France**

Sujets de recherche de fonds pour cette année : criminologie et méthodes statistiques avancées spatio-temporelles + Méthodes statistiques adaptées aux tendances spatiales

Intérêts :

-Zoniers plus intelligents, faire mieux que le simple zonier pauvreté point par point par exemple en regardant les voisins autrement dit : (x,y)=pauvre mais qu’en est-il des zones voisines et quelle est l’influence de la richesse des voisins ?

-Ecosystème : afficher un indicateur de vol associé à une maison et détecter/informer sur une tendance locale (dans le temps et l’espace) au vol ie indicateur de risque de « crise ».

D’où le problème : détecter le risque, valider la certitude.

-Création de lien avec écoles actuariat ou data sciences.

Les ENSAE lisent et approfondissent la théorie sous jacente pour avoir les intuitions et éviter les spurious regressions en cherchant « au pif » à créer du sens. Il faut être un minimum capable d’orienter les méthodes quantitatives pour distinguer les coïncidences pures des vrais sujets. (sauf serendipité … ?)

La théorie derrière doit être solide pour qu’on puisse avoir confiance dans les résultats, premier critère : reproductibilité ?

Ce qui a déjà été fait sur ces données

Modèle actuel : uniquement charge et non fréquence, concernant le vol dans les appartements à Paris.

Choix des coûts parce que fréquence semblait plutôt diffuse (homogène) vs charge très variable en est-ouest.

Tout ce travail a été réalisé dans l’optique plus global d’un renouvellement du zonier vol d’AXA France, objectif non-accompli, on veut aller plus loin que la maille INSEE, faire un vrai découpage par rue.

1 année de données et 1million d’observations… Ca n’est pas énorme.

Problème de fréquence autour de 1% : 10k pour tout paris c’est trop peu.

En parallèle, projet de géolocalisation des risques. Actuellement vision instantanée sur le portefeuille prévue pour la fin d’année mais l’objectif est de remonter à 5 ans.

Mais il y a un sujet autour de l’adresse, on ne peut pas géolocaliser les adresses parce que problème d’entrée des données. 1,8M de courriers par an qui reviennent (6,6M de clients MRH).

Actuellement sous-traité à AGS (très cher !? Mais apparemment ils pallieraient le problème de lecture des adresses…)

Retour sur notre problématique :

Demande d’échantillon : le plus hétérogène possible avec un long historique mais pas trop grand pour qu’on puisse « regarder » graphiquement les choses à une échelle fine (observation par quartier ?).

Choisir une zone avec une grosse fréquence de vol pour que la data soit plus dense.

Idéalement beaucoup d’info mais elle doit être exploitable et riche (prête à l’emploi, cohérente, non redondante). Attention aux changements économiques ou politiques exemple changement de la franchise en 2008…

En pratique :

Partager le travail Base française, base belge, qui fait la lecture de la théorie ? qui s’intéresse plus particulièrement à l’implémentation ? N’oublions pas que les étudiants devront récupérer eux même la donnée de geocoding, ca fait partie de l’intérêt du travail, on leur fournit une matière actuarielle traditionnelle et ils la transforme en « néo-data » pour DSA (data scientist actuary). To be continued.

Idée pour la db : recouper les bases de données MRH souscription avec fiches de client également disponible dans des bases SAS mais le renseignement peut-être problématique, modalités NA et « ? » ou illisibles.

Sorite : Bris de glace à rapprocher du vol ?

DCS service fraude véronique brignolas ?

Problème du travail de GBS et AG :

Pas assez d’orientation du travail à partir de la littérature.

Préférence pour des algorithmes qui « trouvent » tout seul la solution et propose une séparation optimale en intégrant toutes les variables en compétitions.

Pas de modélisation des interactions et pas de sélection des variables.

Problème de stabilité ? Gain négligeable en performance prédictive.

Le type de données extraites est trop homogène (voisins : bar, restau, vétérinaire, hall, gare…).

Il faut traiter au moins l’un des 3V (vélocité, variété, volume).

Random forest c’est juste un bootstrap sur un arbre de décision CART.