Modèle prédicatif du vol à vélo dans la ville de Toronto à partir de séries chronologiques et de données démographiques

Rapport technique

MORENO Tatiana*, CARREÑO Diego* et MEJIA Johan*
* MSc. IT : DSc. - IMT Atlantique
jeny-tatiana.moreno-perea@imt-atlantique.net
diego-andres.carreno-avila@imt-atlantique.net
johan-steven.mejia-mogollon@imt-atlantique.net

I. Résumé

Le présent rapport identifie comme problématique le vol de bicyclettes dans la ville de Toronto, qui est en augmentation, en raison des facteurs de croissance démographique et de l'utilisation de la bicyclette comme moyen de transport. Par conséquent, l'utilisation de différentes architectures de réseaux neuronaux récurrents est proposée pour prédire la densité du pourcentage de vols dans 27 régions, en regroupant les séries chronologiques sur la base de techniques de classification hiérarchique. En conséquence, il a été constaté que la distribution des séries temporelles peut être décrite avec deux architectures de réseaux neuronaux récurrents, basées sur des cellules *Long Short-Term Memory*, ce qui laisse la possibilité d'améliorer l'architecture et de l'étendre à toutes les régions de Toronto.

II. Introduction

Au fil des ans, les gens ont peu à peu peu peuplé la terre, créant des interactions complexes, qui dépendent des besoins du moment. Des études montrent que, de 1950 à 2009, la croissance démographique a été de quatre milliards, et devrait atteindre 9.2 milliards d'ici l'an 2005, soit une croissance prévue de 41 % et une croissance estimée à 75 millions de personnes par an [1]. Cependant, on s'attend à ce que d'ici le nouveau siècle, cette tendance change et que la probabilité de croissance des personnes diminue [2].

Non seulement l'augmentation de la population mondiale a changé, mais les besoins actuels ont également changé. Préoccupation pour les bonnes habitudes alimentaires et l'exercice physique [3], [4], les économies d'énergie et le souci de l'environnement [5], [4], les changements de mentalité dans les aspects culturels et politiques [6], entre autres, montrent que les besoins des gens ont été transformés par le passage du temps.

Plus précisément, l'un des aspects les plus importants en tant que problème actuel est la mobilité et le transport. Selon certaines études, la gestion et l'activité des transports quotidiens est l'une des clés fondamentales de la croissance économique d'un pays, mais entraîne à son tour d'autres problèmes tels que les dégâts climatiques et la forte consommation d'énergie [7]. Le problème de la mobilité dans les villes est si important que même les modèles de transport ont été développés [8] pour en diminuer l'impact [9].

Compte tenu de la croissance démographique de ces dernières années, qui entraîne une augmentation de l'utilisation des transports publics dans les différents territoires de la planète et qui entrave la mobilité des personnes, diverses alternatives ont été trouvées [10], [11]. En particulier, l'utilisation de la bicyclette a considérablement augmenté au cours de la dernière décennie en tant que moyen de transport durable. Aujourd'hui, plus de 800 villes utilisent le vélo comme moyen de transport [12], [13], et l'on estime que d'ici 2050, il pourrait y avoir cinq milliards d'utilisateurs de la bicyclette, plus 50 % de la population mondiale sachant comment faire de la bicyclette [14]. Une des raisons en particulier est la sécurité des blessures. Un exemple en est donné par les villes d'île-de-France et du Grand Londrès, où le nombre de personnes blessées sur des trajets de plusieurs milliards de kilomètres n'est que de 631 pour les personnes possédant leur propre vélo, et de 253 pour les personnes utilisant des vélos partagés, dont seulement 25 et 13 sont des cas mortels, respectivement [12].

Cependant, l'augmentation du nombre de vélos a engendré un nouveau problème au fil des ans. Selon les données recueillies par l'enquête internationale sur les victimes de la criminalité dans différents pays, des tendances de 4.7% sont observées pour les vols à l'encontre des cyclistes, contrairement aux vols à l'encontre des automobilistes et des motards, dont les pourcentages sont respectivement de 1.2 % et 1.9 % [15]. En outre, la même étude montre que seulement 56% des

cas sont signalés, ce qui entraîne une augmentation réelle des vols de vélos.

En particulier, l'une des villes touchées par ce problème est la ville de Toronto, au Canada. En 2018, 3937 vélos ont été volés, et on estime qu'environ 144 vélos sont volés pour 100000 habitants. A partir de 2013, on observe une augmentation moyenne de 215 ± 165 vols de vélos par an, ce qui implique un coût moyen des réclamations des utilisateurs aux assureurs de $2084 dollars entre 2014 et 2018 [16].

Il convient donc de mener une étude pour déterminer le pourcentage de vols sur le plan démographique. En outre, il serait utile de disposer d'un modèle qui permette d'utiliser les registres actuels des vols de vélos pour prévoir le nombre de vols qui se produiront dans les semaines à venir, afin d'aider les entités concernées à réduire ce problème grâce à des campagnes et de la publicité, des stratégies visant à améliorer la sécurité publique et d'autres outils que la ville de Toronto autoriserait. C'est à ce stade que la science des données fournit certains outils qui permettent de prendre des décisions en dépit de certains inconvénients, "à condition que l'on adopte une attitude critique et que tant le choix des données que leur analyse s'inscrivent dans des contextes historiques et sociaux" [17]. Diverses applications dans le domaine de la médecine [18], [19], des finances [20], des systèmes de recommandation [21], des les informatiques [22], [23], [24], [25] et de l'énergie [26], [27], ont été développées dans les domaines de la connaissance de l'apprentissage automatique et de la science des données.

Par conséquent, ce document développe une analyse des conditions technologiques, géographiques et opérationnelles pour l'application d'un modèle d'analyse de données qui contribuera à réduire le taux de vol de vélos dans la ville de Toronto, sur la base de la prédiction hebdomadaire de la densité de vol par surface entre les années 2014 et 2019.

Le rapport est décrit comme suit : la section III présente la base de données utilisée dans la recherche. Cette base de données est analysée dans la section IV. Ensuite, le modèle de solution est présenté dans la section V, et les résultats et conclusions sont analysés dans la section VI et section VII, respectivement.

III. Description de l'information

Deux bases de données sont disponibles pour le développement de la recherche, concernant les données démographiques et les enregistrements des vols de vélos dans la ville de Toronto. Elles sont décrites dans les sections sous-section III-A et sous-section III-B, pour enfin expliquer la procédure de création de la base de données composée des deux (sous-section III-C).

Il est précisé que les deux bases de données sont à code source ouvert, sous réserve que Statistique Canada accorde une licence mondiale, libre de redevance et non exclusive. Il est de la plus haute importance de préserver l'intégrité et la vie privée des personnes. Par conséquent, aucun résultat de ce rapport n'est basé sur des informations privées ou individuelles de particuliers. Les deux bases de données sont référencées avec leurs liens de téléchargement respectifs, qui contiennent toutes les informations des auteurs, afin de sauvegarder les droits de propriété intellectuelle.

A. *Base de données sur la population*

Des informations sur les différentes caractéristiques de la population de Toronto sont disponibles grâce à l'application développée par [28], appelé *Canada census* (`cancensus`) (Figure 1).

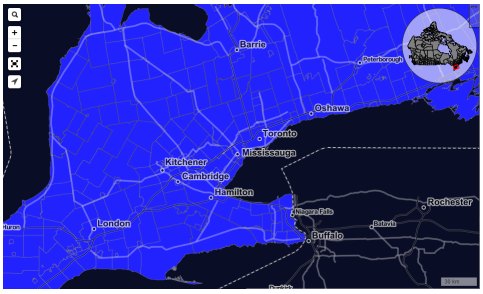

Figure 1. Exemple d'utilisation de l'application `cancensus`

L'application `cancensus` est un ensemble de différentes bases de données géographiques socio-économiques pour la ville du Canada. Il est possible d'y connaître différents aspects à différents niveaux d'agrégation, tels que les provinces, les districts, les zones métropolitaines, les subdivisions, les zones de surveillance et de diffusion. De plus, l'outil permet [28] :

- Télécharger les données et la géographie du recensement dans un format propre et prêt pour l'analyse
- Des outils pratiques pour rechercher et travailler avec les régions de recensement et les hiérarchies de variables
- Fournit la géographie du recensement en plusieurs formats spatiaux R
- Fournit des données pour les recensements de 2016, 2011, 2006, 2001 et 1996
- Accès aux données des contribuables au niveau du secteur de recensement pour les années fiscales 2000 à 2017.

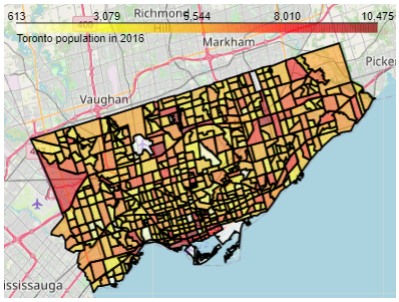

Figure 2. Représentation géospatiale de la population de Toronto d'ici 2016 par CT.

Pour ce rapport, une étude sera réalisée au niveau de l'agrégation de suivi, appelé CT, car à ce niveau, les informations sont regroupées de telle sorte qu'elles permettent une analyse concrète (pas si générique) des informations, sans atteindre le cas extrême de ne pas présenter les enregistrements de vols de vélos (comme cela se produit à d'autres niveaux d'agrégation plus détaillés). Afin de voir la structure des informations, la Figure 2 montre les divisions de la ville de Toronto en utilisant le niveau d'agrégation de CT. Certaines des données pertinentes pour la recherche sont présentées dans la description du Tableau I. Au total, il y a 572 divisions dans la catégorie des CT, dont nous disposons d'informations provenant de 475 enregistrements dans le champ `v_CA16_5807: Bycicle`, une propriété importante dans le développement du modèle.

Tableau I
PRÉSENTATION DE QUELQUES CHAMPS IMPORTANTS OBTENUS AVEC L'API CANCENSUS.

Champ	Description du champ	Type	μ	σ	min()	max()
GeoUID	ID de la région	String	—	—	—	—
Area (Km^2)	Superficie en Kilomètres carrés	float	1.1091	1.4264	13×10^{-3}	20.225
Population	Nombre de personnes dans la région	int	4784	1978	64	17549
Dwellings	Nombre de maisons	int	2068	1217	19	15207
Households	Nombre de personnes qui vivent dans la maison	int	1952	1069	19	11891
v_CA16_406	Densité de population par kilomètre carré	float	7853	8055	13.8	82433.8
v_CA16_5807	Vélos : Nombre total de vélos dans une région	int	79	107	10	740

D'autre part, l'identifiant de chaque piste de recensement est contenu dans le champ GeoUID.

B. Base de données sur les vols de vélos

Les services de police de Toronto fournissent une base de données open source des dossiers de vols de vélos dans toute la ville de 2014 à 2019. Il s'agit d'une base de données qui assure la confidentialité des informations, de sorte que les données sont biaisées par rapport aux enregistrements réels, et aucune information personnelle n'est disponible pour chaque individu, ce qui n'a pas beaucoup d'impact sur l'objectif du projet. Les dossiers peuvent être téléchargés sur le site officiel des services de police de Toronto. La base de données contient principalement celles décrites dans le Tableau II. La latitude et la longitude fournissent la paire de coordonnées permettant de localiser le vol, en liaison avec la date décrite par le champ `Ocurrence_Date`. Enfin, le coût et le statut de l'enregistrement (volé, récupéré ou inconnu) sont expliqués respectivement dans les champs `Cost_of_Bike` et `Status`.

Tableau II
PRÉSENTATION DE QUELQUES CHAMPS IMPORTANTS DANS LA BASE DE DONNÉES DES VOLS DE VÉLOS FOURNIE PAR LE SERVICE DE POLICE DE TORONTO.

Champ	Description du champ	Type	Éléments uniques
Lat	Latitude du dossier (coordonnée géopositionnelle)	float	4874
Long	Longueur du dossier (coordonnée géopositionnelle)	float	4885
Occurrence_Date	Date of occurrence	Timestamp	2104
Cost_of_Bike	Cost of Bicycle	float	1458
Status	Status of event	String	3

C. Liaison des bases de données

La latitude et la longitude étaient les champs clés pour relier les bases de données. La CT de chaque enregistrement est estimée sur la base de sa position géopositionnelle. Une fois que chaque enregistrement est lié à son identifiant correspondant (GeoUID), les données du recensement peuvent être utilisées dans chaque enregistrement de vol. Il convient de noter que le niveau d'agrégation des informations a été effectué par semaine et par zone géographique (CT), ce qui a donné lieu à la possibilité d'estimer différents paramètres pour chaque niveau d'agrégation.

En principe, le nombre total de vélos volés (`theft_bikes`) est défini comme le nombre d'enregistrements de vélos volés qui sont encore volés, c'est-à-dire dont le `status` est égal à volé. Par conséquent, l'Équation 1 présente la définition du calcul de la densité du pourcentage de vols de bicyclettes par surface (DBTS) et pour chaque niveau de regroupement, produit de l'estimation du pourcentage de vols par zone.

$$\texttt{DBTS} = 100 \times \frac{\texttt{theft\_bikes}}{\texttt{v\_CA16\_5807}} \times \frac{1}{\texttt{Area}} \tag{1}$$

Si nous examinons le résultat de l'estimation DBTS, nous verrons différentes séries chronologiques correspondant à chaque CT (Figure 3).

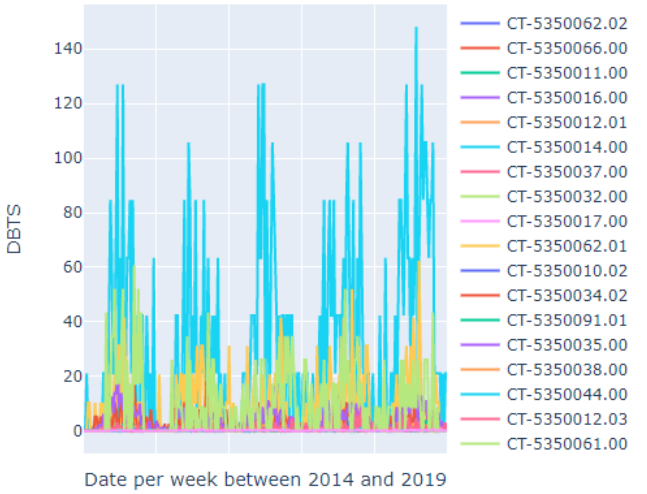

FIGURE 3. Séries chronologiques par CT et par semaine de la DBTS

Les graphiques de la Figure 3 ont été réalisés en supposant comme vol nul les niveaux d'agrégation dont les enregistrements étaient inexistants. On suppose alors qu'au cours d'une semaine donnée sans signalement, il n'y a pas eu de vols de vélos, soit par désinformation, soit par manque d'enregistrement. Cette procédure a été réalisée dans le but de ne pas biaiser les bases de données, en créant des informations non vérifiées.

IV. ANALYSE DES DONNÉES

Dans cette section, nous présentons quelques statistiques réalisées dans l'ensemble de données unifiées, dans le but

de comprendre son comportement et de définir les principaux paramètres de l'architecture à proposer.

A. *Distribution des données*

La Figure 4 montre la répartition des données dans les trois catégories de la page à onglet `status`.

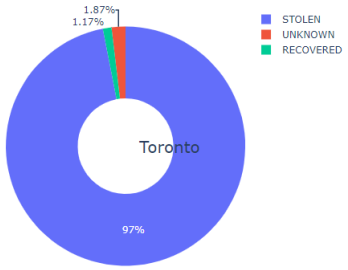

FIGURE 4. Distribution du label `status` dans l'ensemble de la base de données.

Nous pouvons voir que la plupart des données sont concentrées dans la catégorie `STOLEN`, ce qui nous permet de tirer profit de 97% des informations avec le nombre de records pour l'estimation du nombre de vélos volés. 3% des données seront rejetées.

Nous définissons le nombre de vols comme la somme du nombre total d'enregistrements ayant la catégorie `STOLEN` pour les différents niveaux d'agrégation, et le pourcentage de vélos volés comme le nombre de vols pour le nombre total d'enregistrements de vélos dans chaque région. Il convient de préciser que le nombre total de vélos par région est une statistique tirée de la base de données `cancensus`, de sorte qu'il n'y a qu'un enregistrement tous les cinq ans. On suppose que le nombre de vélos reste constant pendant ces périodes.

D'autre part, la Figure 5 montre la densité au fil des ans. Nous voyons comment, depuis 2014, ce problème s'aggrave jusqu'en 2018, pour diminuer l'année dernière (peut-être en raison du manque de données).

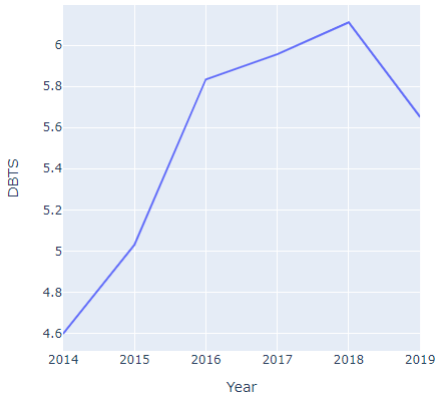

FIGURE 5. Évolution de la densité du pourcentage de vols par surface dans la ville de Toronto pour les années 2014 à 2019.

En outre, nous souhaitons présenter la Figure 6 comme le résultat de l'analyse des informations par mois et par jours. Le graphique montre que le plus grand nombre de vols est effectué dans les derniers mois de chaque année pour la ville de Toronto.

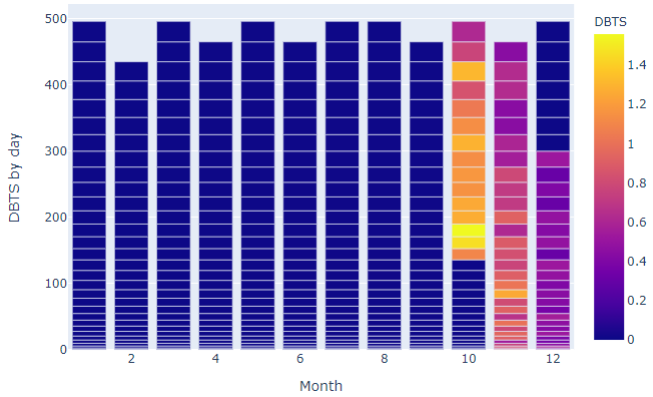

FIGURE 6. Carte de chaleur du DBTS par jour.

Enfin, une analyse détaillée de la relation entre le `DBTS` à travers les différents jours de la semaine de chaque mois de l'année est présentée, afin d'estimer la probabilité par jour de la semaine que le vol aurait été effectué, en fonction du mois de l'année.

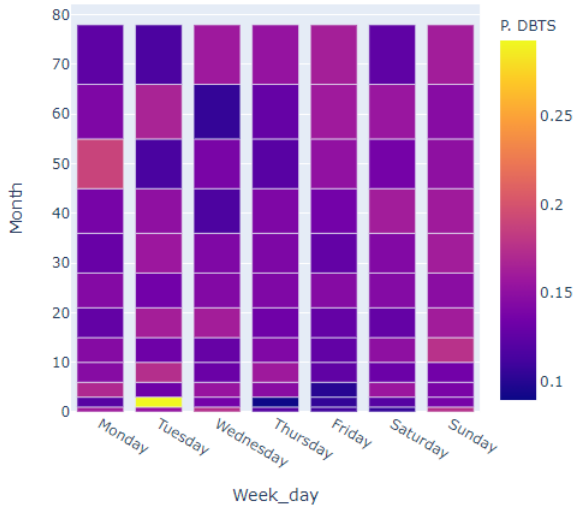

FIGURE 7. Probabilité de `DBTS` par jour de la semaine en fonction du mois de l'année pour la ville de Toronto.

En général, on constate que le plus grand nombre de vols se produisent les premiers jours de la semaine, avec une plus grande incidence le mardi en février.

V. DÉFINITION DE L'ARCHITECTURE

Après avoir présenté le contexte général du problème et l'introduction de la base de données, il a été décidé d'expliquer en profondeur l'objectif de ce document, compte tenu de la préoccupation concernant l'augmentation des vols de bicyclettes ces dernières années (Figure 5).

Un modèle de prédiction des séries chronologiques est nécessaire pour estimer le DBTS par CT et par date, pour la ville de Toronto. A cette fin, les séries chronologiques présentées dans la Figure 3 serviront de base. Aspects importants à prendre en compte :

— L'échelle de temps à analyser sera celle des jours. Toutefois, étant donné la quantité limitée de données à ce niveau de spécificité, les prédictions seront faites au niveau de la semaine, et la probabilité de densité de vols par jour sera fournie sur la base des résultats de la Figure 7

— Les dossiers inexistants sur les vols de vélos seront considérés comme nuls, ce qui réduira le niveau de partialité des informations

— Les informations des CT sans valeur dans le champ v_CA16_5807, nécessaires pour estimer la DBTS (Équation 1), seront rejetées.

Dans cette optique, nous présentons la procédure de regroupement hiérarchique des séries temporelles dans la sous-section V-A, qui sera utilisée pour le développement du réseau neuronal récursif (sous-section V-B).

A. *Regroupement hiérarchique*

Le clustering est une technique d'exploration de données permettant de regrouper un ensemble d'objets de telle sorte que les objets d'un même groupe soient plus semblables les uns aux autres qu'à ceux d'autres groupes.

Dans la classification hiérarchique, chaque objet (point de données) est affecté à une grappe distincte. Ensuite, nous calculons la distance (métrique de similarité) entre chacune des grappes et nous rejoignons les deux grappes les plus similaires.

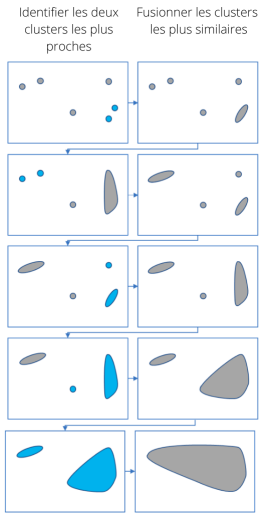

FIGURE 8. Clustering hiérarchique

La mise en grappes hiérarchique commence par traiter chaque observation comme une grappe distincte. Ensuite, elle exécute les deux étapes suivantes de manière répétée : (1) identifier les deux grappes les plus proches l'une de l'autre, et (2) fusionner les deux grappes les plus similaires. Cette opération se poursuit jusqu'à ce que tous les groupes soient fusionnés ensemble.

Le principal résultat de la classification hiérarchique est un dendrogramme, qui montre la relation hiérarchique entre les groupes. Un dendrogramme est un diagramme en forme d'arbre qui enregistre les séquences de fusions ou de divisions.

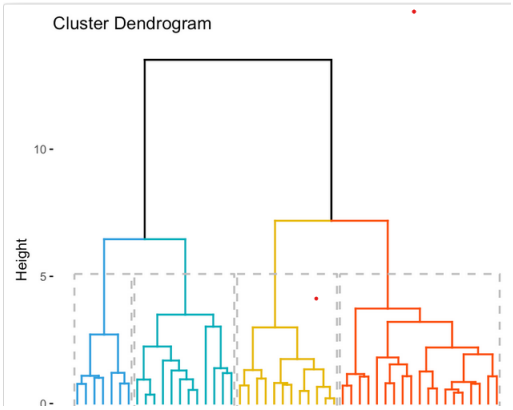

FIGURE 9. Dendrogramme de classification hiérarchique

Le calcul de la similarité entre deux clusters est important pour fusionner les clusters. Dans notre projet, nous utilisons la méthode ward's, la façon de calculer la similarité entre deux groupes est de prendre toutes les paires de points et de calculer leurs similarités et de calculer la moyenne des similarités, puis de calculer la somme du carré des distances entre deux points qui appartient à des groupes différents. L'expression mathématique est la suivante :

$$\text{similarity}(C1, C2) = \frac{\sum(\text{dist}(P_i, P_j))^2}{|C_1| * |C_2|} \tag{2}$$

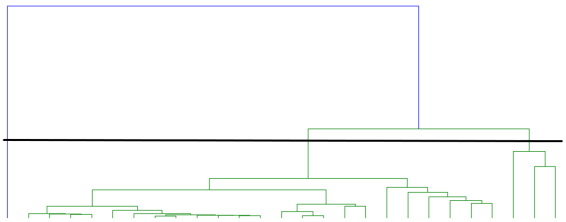

FIGURE 10. Dendogramme de classification hiérarchique regroupé par CT

Un algorithme de regroupement est nécessaire pour regrouper les séries temporelles appartenant aux CT ayant des tendances similaires afin de faire une prévision plus efficace. À cette fin, le modèle de regroupement hiérarchique sera utilisé, en prenant comme indices les séries chronologiques présentées dans la Figure 3. Comme le montre la Figure 10 , le troisième niveau du modèle a été retenu (k = 3), les 26 CT les plus pertinents et les plus documentés ont été regroupés en 3 groupes afin de prévoir le comportement des vols dans

les régions où il n'y a pas de données ou trop de lacunes. Parmi ces trois groupes, seuls les groupes comportant le plus grand nombre d'éléments seront analysés (un des groupes ne contient qu'un seul CT).

Le regroupement des CT par groupes est le suivant :

— Cluster 1 :
[5350008.01 5350008.02 5350010.02
5350012.01 5350012.03 5350012.04
5350013.02 5350016.00 5350017.00
5350034.02 5350035.00 5350037.00
5350044.00 5350062.02 5350064.00
5350089.00 5350091.01 5350092.00
5350011.00 5350013.01 5350032.00
5350066.00 5350038.00]
— Cluster 2 : [5350015.00, 5350061.00, 5350062.01]

B. *Réseau neuronal récurrent*

Afin de faire la prédiction du pourcentage de vols par km^2 dans les CT de la ville de Toronto par semaine, un réseau Long Short-Term Memory (LSTM) récurrent est fait qui compte avec un certain nombre de blocs empilés et un nombre de couches cachées. La Figure 11 présente un module LSTM classique.

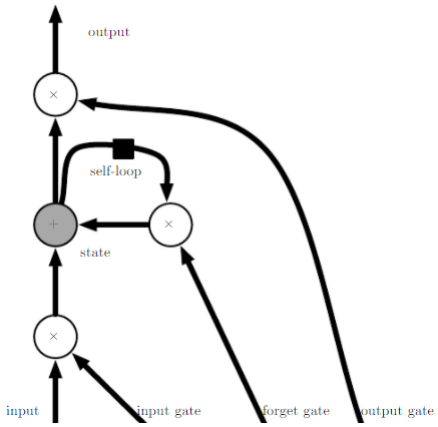

FIGURE 11. Description détaillée du réseau représentant une cellule LSTM [29].

De plus, l'Équation 3 explique la fonction qui décrit le résultat des cellules LSTM.

$$\mathbf{Q}^{(t)} = \sigma\left(b^{O} + \mathbf{U}^{O}_{hx}\mathbf{x}^{(t)} + \mathbf{W}^{O}_{hh}\mathbf{H}^{(t-1)}\right) \tag{3}$$

Avec $\mathbf{Q}^{(t)}$ la sortie pour l'état t, b le vecteur de biais, $\mathbf{U}_{hx}/\mathbf{W}_{hh}$ les matrices de poids pour l'entrée $\mathbf{x}^{(t)}$ et $\mathbf{H}^{(t-1)}$ (états internes en l'état $t-1$), respectivement. Contrairement à une Recurrent Neural Network (RNN) classique, une cellule LSTM possède un *forget gate* qui réduit le problème de la disparition du gradient (*vanishing gradient*). Cela modifie l'équation des états internes l'ensemble des équations décrites dans Équation 4 [29].

$$\begin{aligned}
\mathbf{F}^{(t)} &= \sigma\left(b^{f} + \mathbf{U}^{f}_{hx}\mathbf{x}^{(t)} + \mathbf{W}^{f}_{hh}\mathbf{H}^{(t-1)}\right)\\
\mathbf{S}^{(t)} &= \mathbf{F}^{(t)}\mathbf{S}^{(t-1)} + \mathbf{G}^{(t)}\sigma\left(b + \mathbf{U}_{hx}\mathbf{x}^{(t)} + \mathbf{W}_{hh}\mathbf{H}^{(t-1)}\right)\\
\mathbf{G}^{(t)} &= \sigma\left(b^{g} + \mathbf{U}^{g}_{hx}\mathbf{x}^{(t)} + \mathbf{W}^{g}_{hh}\mathbf{H}^{(t-1)}\right)\\
\mathbf{H}^{(t)} &= \tanh\left(\mathbf{S}^{(t)}\right)\mathbf{Q}^{(t)}
\end{aligned} \tag{4}$$

En tenant compte de la classification des CT ayant des comportements similaires obtenue dans la section précédente, 2 modèles sont développés pour obtenir la prédiction de l'ensemble de la ville. Le code développé se trouve dans les annexes . La Figure 12 montre le schéma fonctionnel du réseau mis en œuvre . Pour chaque grappe, une configuration du réseau LSTM est réalisée, en faisant varier les hyper paramètres du modèle.

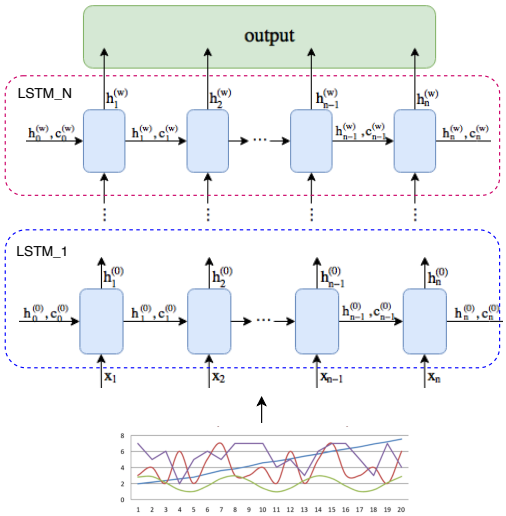

FIGURE 12. Réseau récurrent LSTM

Pour faire la prédiction, comme entraînement l'information historique des vols de 2014 à 2018, et le test /validation sur l'information sur le vol des vélos de l'année 2019 est fait. La série de formation est composée de la série chronologique de semaines du 14-04-2014 au 31-12-2018, la série de validation est la série de semaines du 01-01-2019, du 30-06-2019 et la série de test est du 01-07-2019 au 31-12-2019.

Tableau III
SÉRIE D'EXPÉRIMENTATIONS

Cluster	hidden dim	rnn layers	dropout	Lr	epochs	Best loss
1	300	4	0.3	1e-5	200	**14.47**
1	500	4	0.4	1e-5	300	15.89
1	500	2	0.3	1e-4	300	14.72
2	300	4	0.3	1e-5	200	34.179
2	700	7	0.8	1e-4	500	32.89
2	600	4	0.8	1e-4	200	**30.02**

Plusieurs tests sont effectués afin de déterminer le meilleur modèle pour chaque cluster. Lors de la formation du cluster,

celui qui présente une erreur de validation mineure est enregistré comme le meilleur modèle. Le Tableau III présente certains des meilleurs modèles obtenus par grappe.

La section VI présente les résultats obtenus lors de la formation de chaque modèle.

VI. Analyse des résultats

Cette section présente les résultats des modèles prédictifs réalisés pour les groupes de clusters présentés précédemment.

A. *Cluster 1*

Dans le graphique de la Figure 13 la série chronologique de certaines CTs correspondant au cluster 1 est présentée. Ces CTs ont la caractéristique que la série chronologique contient plus d'informations, de sorte que la tendance des vols par semaine dans la ville est plus clairement présentée.

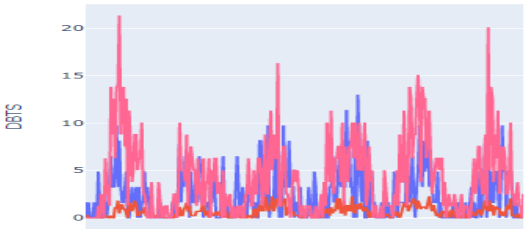

Figure 13. Série temporelle CTs cluster 1

La Figure 14 montre les pertes de formation et de validation obtenues pour le meilleur modèle avec configuration selon le Tableau III. Le critère utilisé dans la formation était la somme de l'erreur quadratique moyenne, étant donné que les valeurs (une fois normalisées) présentent un Mean Square Error (MSE) moyen proche de l'ordre 1×10^{-3}, ce qui représente une mauvaise prédiction de valeurs apparemment acceptables. Il est admis que l'erreur de validation aura tendance à être plus faible puisque les ensembles de validation contiennent moins de données que l'ensemble de formation. Afin d'être plus rigoureux dans la formation, l'opération moyenne a été remplacée par la somme. Une valeur de validation minimale de 14.47 est obtenue.

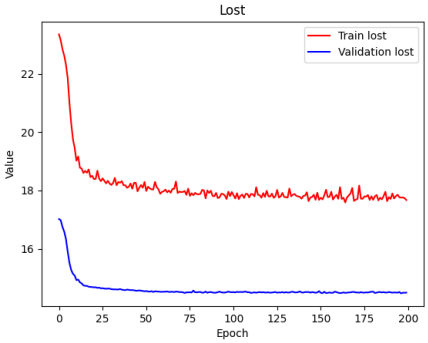

Figure 14. Lost per epoch

Dans la Figure 15, la valeur prédite (blue) en fonction de l'ensemble de formation est présentée, on observe que la prédiction est basée sur la valeur moyenne des vols. Une valeur de 9.409 est trouvée pour MSE.

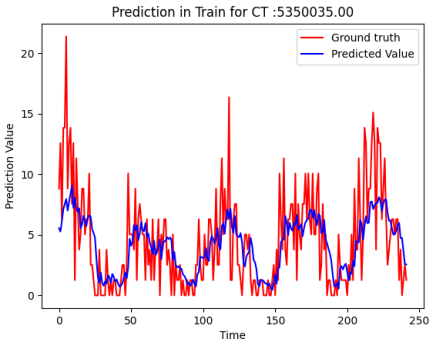

Figure 15. Prédiction dans train data set MSE = 9.409

La Figure 16 montre la valeur prédite par rapport à la valeur réelle obtenue dans l'ensemble de validation pour une CT, nous avons un MSE de 4,23639. On peut voir que la prévision suit la tendance générale du pourcentage de vols

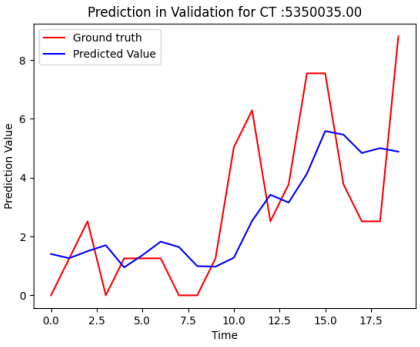

Figure 16. Prédiction dans validation data set MSE = 4.23639

La Figure 17 montre la valeur prédite par rapport à la valeur réelle obtenue dans l'ensemble de validation pour une CT, nous avons un MSE de 13.236. On peut voir que la prévision suit la tendance générale du pourcentage de vols

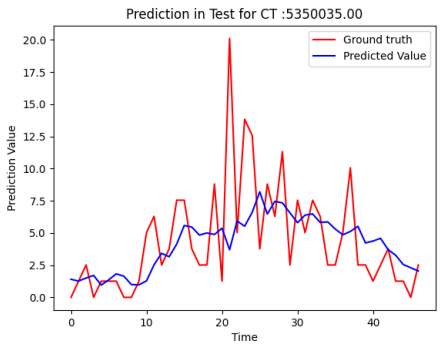

Figure 17. Prédiction dans test data set MSE= 13.236

B. Cluster 2

Dans le graphique de la Figure 18 la série chronologique de certaines CTs correspondant au cluster 2 est présentée. Ces CTs ont la caractéristique qu'ils présentent un grand nombre d'informations manquantes, ce qui rend difficile une prévision correcte pour l'année 2019.

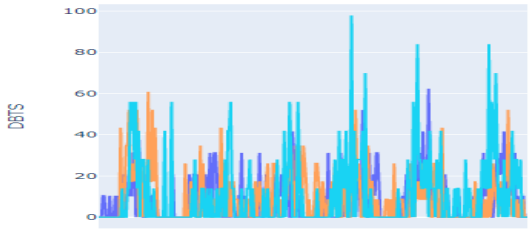

FIGURE 18. Série temporelle CTs cluster 2

La Figure 19 montre les pertes de formation et de validation obtenues pour les CTs avec la configuration selon le Tableau III. Une valeur de validation minimale de 30.02 est obtenue.

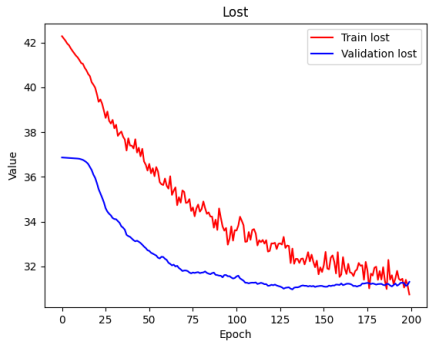

FIGURE 19. Lost per epoch

Dans le graphique de la Figure 20, la valeur prédite (blue) en fonction de l'ensemble de formation est présentée, on observe que la prédiction est basée sur la valeur moyenne des vols. Une valeur de 108.54112 est trouvée pour MSE.

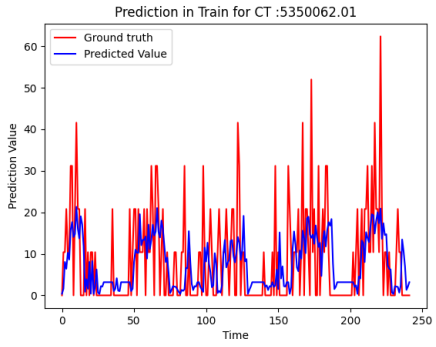

FIGURE 20. Prédiction dans train data set MSE : 108.54112

La Figure 21 montre la valeur prédite par rapport à la valeur réelle obtenue dans l'ensemble de validation pour une CT, nous avons un MSE de 60,6104. On peut voir que la prévision suit la tendance générale du pourcentage de vols

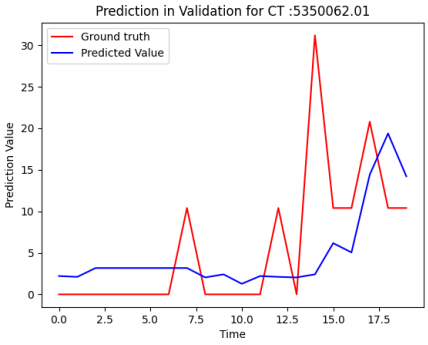

FIGURE 21. Prédiction dans validation data set MSE = 60.6104

La Figure 22 montre la valeur prédite par rapport à la valeur réelle obtenue dans l'ensemble de validation pour une CT, nous avons un MSE de 114.4189. On observe que la prédiction ne permet pas de détecter les changements dans le pourcentage de vols entre les semaines

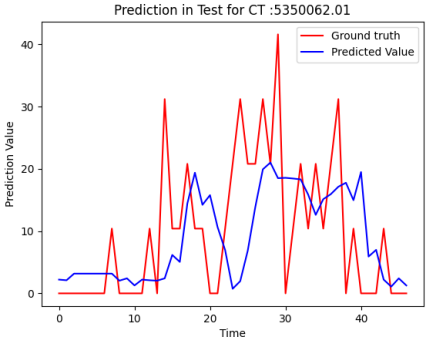

FIGURE 22. Prédiction dans test data set MSE= 114.4189

VII. CONCLUSIONS

Il existe de nombreuses applications dans lesquelles la science des données peut être très utile. Dans ce travail, nous avons développé un modèle qui pourrait être utilisé par les entités gouvernementales pour aider à prévenir et à réduire le vol de vélos dans la ville de Toronto.

Afin que la prévision soit aussi précise et sans erreur que possible, il a été décidé de mettre en œuvre un algorithme de regroupement hiérarchique qui regroupe les séries temporelles des CT les plus informatives, et il a été observé que cet algorithme était capable de trouver les similitudes intrinsèques des CT dans la base de données. Pour constater que les CT du cluster 2 ont un comportement de vol similaire dans le temps.

Pour plus de rigueur, le critère de MSE sum a été choisi. Bien qu'elle ne fournisse pas d'informations comparables entre

les deux processus (formation et validation), elle nécessite une formation plus ciblée pour trouver une solution dans laquelle, pour chaque point de l'ensemble de formation, on obtient le minimum d'erreur possible. Toutefois, les résultats montrent que la série prédite tend vers la moyenne de la série originale, ce qui prouve que le réseau permet de rechercher la distribution mondiale de l'information. Il est possible de diminuer l'erreur d'apprentissage en augmentant les couches `LSTM` ou les états internes, ou en utilisant d'autres techniques qui se concentrent sur la diminution du biais de l'architecture. Plusieurs architectures de réseau récurrentes seront testées à l'avenir.

En général, on observe que le manque d'information affecte dans une large mesure l'obtention d'un bon modèle. Le modèle du cluster 1 présente une meilleure performance que celui proposé pour la cluster 2, néanmoins les deux possèdent des erreurs élevées au moment de prédire le pourcentage de vols.

D'après les résultats obtenus avec les modèles mis en œuvre, il est possible de prévoir le pourcentage moyen de vols par zone dans les différents CT, mais pour les travaux futurs, afin d'affiner le modèle, davantage d'informations sont nécessaires.

Enfin, il est prévu de développer à l'avenir une application qui permettra de voir l'évolution du `DBTS` par région sur une base quotidienne, ainsi que d'autres statistiques importantes. En outre, il permettrait de regrouper les informations par zones (comme le fait actuellement le `cancensus`) et par jours, à des échelles spatio-temporelles (Figure 23).

VIII. Annexes

Le code, ainsi que la manipulation de la base de données, sont contenus dans le dépôt bikes-theft-model.Dans le dépôt, vous trouverez comme code descriptif le fichier Statistic_bike_theft.ipynb, carnet qui montre en détail toutes les statistiques analysées dans le cadre du développement du projet. En outre, il existe un fichier main_Proyect.py, qui développe la formation à l'architecture `RNN`.

En outre, les spécifications contractuelles des points à prendre en compte pour la signature du projet sont jointes.

Références

[1] J. Bongaarts, "Human population growth and the demographic transition," *Philosophical Transactions of the Royal Society B : Biological Sciences*, 2009.

[2] W. Lutz, W. Sanderson, and S. Scherbov, "The end of world population growth," *Nature*, 2001.

[3] C. university, *The Science Behind Behavior Change*, 2018 (entré le décembre 09, 2020). [Online]. Available : https://www.cuimc.columbia.edu/news/science-behind-behavior-change

[4] N. H. Behaviour, "What works for behaviour change ?" *Nature Human Behaviour*, vol. 2, no. 10, pp. 709–709, Oct 2018. [Online]. Available : https://doi.org/10.1038/s41562-018-0459-4

[5] S. Clayton, P. Devine-Wright, P. C. Stern, L. Whitmarsh, A. Carrico, L. Steg, J. Swim, and M. Bonnes, "Psychological research and global climate change," 2015.

[6] R. Inglehart and C. Welzel, *Modernization, cultural change, and democracy : The human development sequence*. Cambridge University Press, 2005.

[7] Intergovernmental Panel on Climate Change and Intergovernmental Panel on Climate Change, "Transport and its infrastructure," in *Climate Change 2007*, 2012.

[8] M. Wegener, "Overview of Land Use Transport Models," in *Handbook of Transport Geography and Spatial Systems*, 2004.

[9] L. Chapman, "Transport and climate change : a review," *Journal of Transport Geography*, 2007.

[10] I. Spectrum, *Alternative Transportation*, 2020 (entré le décembre 10, 2020). [Online]. Available : https://spectrum.ieee.org/transportation/alternative-transportation

[11] Altenergymap, *Alternative Transportation Articles, Stories & News*, 2020 (entré le décembre 10, 2020). [Online]. Available : https://www.altenergymag.com/tag/alternative-transportation

[12] E. Fishman and P. Schepers, "Global bike share : What the data tells us about road safety," *Journal of Safety Research*, 2016.

[13] B. C. Langford, J. Chen, and C. R. Cherry, "Risky riding : Naturalistic methods comparing safety behavior from conventional bicycle riders and electric bike riders," *Accident Analysis and Prevention*, 2015.

[14] B. Radar, *27 great benefits of cycling*, 2020 (entré le décembre 10, 2020). [Online]. Available : https://www.bikeradar.com/advice/fitness-and-training/30-great-benefits-of-cycling/

[15] A. Sidebottom and S. D. Johnson, "Bicycle Theft," in *Encyclopedia of Criminology and Criminal Justice*, 2014.

[16] K. Chan, *The number of reported bike thefts in Toronto is rising : report*, 2019 (entré le décembre 10, 2020). [Online]. Available : https://dailyhive.com/toronto/toronto-bike-thefts-statistics

[17] L. Daniel and P. N. Ricardo, "Ciencia de datos y estudios globales : aportaciones y desafíos metodológicos," *Colombia Internacional*, pp. 41 – 62, 04 2020. [Online]. Available : http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-56122020000200041&nrm=iso

[18] A. R. Khan, K. T. Hasan, T. Islam, and S. Khan, "Forecasting respiratory tract infection episodes from prescription data for healthcare service planning," *International Journal of Data Science and Analytics*, 2020.

[19] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures," *Biomedical Engineering Letters*, 2018.

[20] J. D. Warnke-Sommer and F. E. Damann, "An improved machine learning application for the integration of record systems for missing US service members," *International Journal of Data Science and Analytics*, 2020.

[21] A. Lysenko, E. Shikov, and K. Bochenina, "Combination of individual and group patterns for time-sensitive purchase recommendation," *International Journal of Data Science and Analytics*, 2020.

[22] N. Levy, A. Rubin, and E. Yom-Tov, "Modeling infection methods of computer malware in the presence of vaccinations using epidemiological models : an analysis of real-world data," *International Journal of Data Science and Analytics*, 2020.

[23] A. Galicia, J. F. Torres, F. Martínez-Álvarez, and A. Troncoso, "A novel spark-based multi-step forecasting algorithm for big data time series," *Information Sciences*, 2018.

[24] A. Ed-daoudy and K. Maalmi, "A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment," *Journal of Big Data*, 2019.

[25] C. Misra, S. Bhattacharya, and S. K. Ghosh, "A fast scalable distributed kriging algorithm using spark framework," *International Journal of Data Science and Analytics*, vol. 10, no. 3, pp. 249–264, Sep 2020. [Online]. Available : https://doi.org/10.1007/s41060-020-00215-3

[26] K. Mulrennan, M. Awad, J. Donovan, R. Macpherson, and D. Tormey, "Modelling the electrical energy profile of a batch manufacturing pharmaceutical facility," *International Journal of Data Science and Analytics*, 2020.

[27] K. Mulrennan, J. Donovan, D. Tormey, and R. Macpherson, "A data science approach to modelling a manufacturing facility's electrical energy profile from plant production data," in *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 2019.

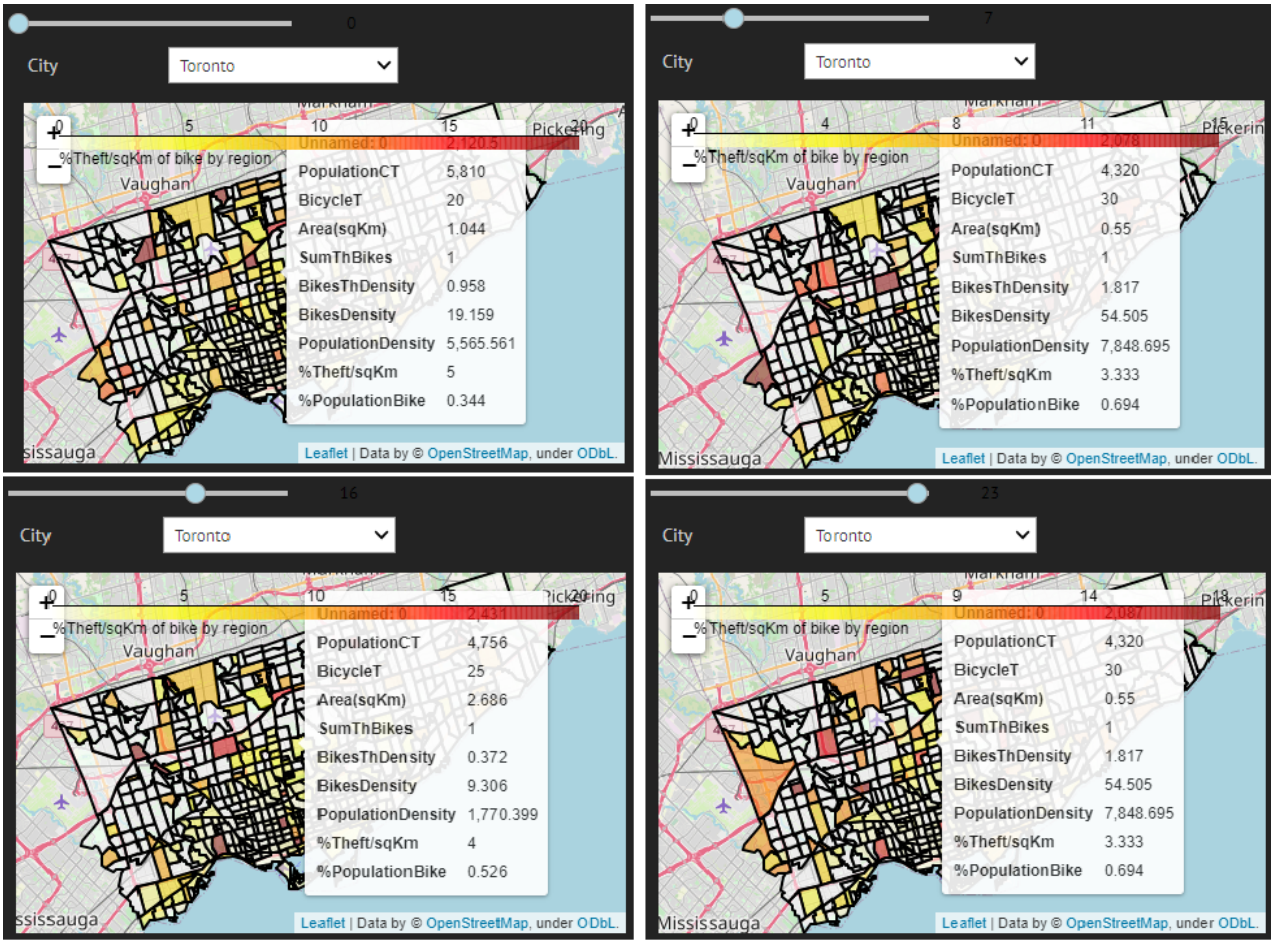

FIGURE 23. Résultats attendus avec l'information de la prédiction des vols de vélos par le CT et les statistiques de population.

[28] J. von Bergmann, D. Shkolnik, and A. Jacobs, *cancensus : R package to access, retrieve, and work With Canadian Census data and geography*, 2020, r package version 0.3.2. [Online]. Available : https://mountainmath.github.io/cancensus/

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.