HEC MONTRÉAL

COVID-19, comportements et sécurité routière : Analyses spatio-temporelles des accidents de la route au Québec

par

Edgar Lanoue

Aurélie Labbe
HEC Montréal
Directrice de recherche

Nicolas Saunier
Polytechnique Montréal
Directeur de recherche

Sciences de la gestion
(Spécialisation Science des données et analytique d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Janvier 2025

Résumé

La pandémie de la COVID-19 a eu de profondes répercussions sur la société québécoise, notamment sur la sécurité routière. On observe effectivement une augmentation relative des accidents mortels depuis le début de la pandémie alors qu'on observe moins d'accidents graves et légers. Il est donc essentiel de comprendre les facteurs contribuant à cette augmentation d'accidents mortels afin de mieux cibler les politiques et interventions en matière de sécurité routière.

L'objectif de ce projet est d'analyser les données d'accidents de 2012 à 2022 pour identifier les phénomènes ou comportements ayant un impact significatif sur les accidents, et ce à différents niveaux : les accidents eux mêmes (analyse ponctuelle) et le nombre d'accidents par municipalité et par MRC. Les principales hypothèses pour expliquer l'augmentation des accidents mortels incluent une hausse des accidents impliquant des usagers vulnérables et une augmentation des comportements risqués sur les routes.

Pour ce faire, nous avons employé plusieurs méthodes d'analyse spatiale, notamment les modèles Besag-York-Mollié (BYM) pour analyser les nombres d'accidents par municipalités et MRC, ainsi que les modèles *Log-Gaussian Cox Process* (LGCP) pour modéliser les emplacements précis des accidents. Ces modèles intègrent plusieurs covariables démographiques et socio-économiques à différents niveaux géographiques, ainsi que des caractéristiques liées aux accidents eux-mêmes.

Cependant, le nombre limité d'accidents mortels, s'élevant à moins de 450 par années à la grandeur du Québec et à à peine une trentaine à Montréal (excluant ceux ayant lieu sur les autoroutes), rend difficile l'obtention de conclusions statistiquement significatives.

En segmentant ces moins de 450 accidents annuels par municipalités ou MRC, on obtient souvent des échantillons très restreints, ce qui complique l'analyse d'autres facteurs en lien avec l'accident comme les distractions, car il faut segmenter encore plus les données. Les modèles ponctuels, quant à eux, se sont révélés très sensibles aux emplacements précis lorsqu'il y a peu de données. En conséquence, il a été difficile d'obtenir des résultats robustes. Malgré ces défis, les analyses ont permis d'identifier des tendances intéressantes et des facteurs potentiellement influents sur l'augmentation des accidents mortels, offrant ainsi des pistes prometteuses pour de futures recherches et pour l'élaboration de politiques de sécurité routière ciblées.

Mots-clés

Analyse spatiale, analyse spatio-temporelle, sécurité routière, modèle BYM, modèle LGCP, langage de programmation *R*, INLA.

Méthodes de recherche

Analyse multivariée, économétrie, recherche quantitative

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

AIA Association des industries de l'automobile du Canada

AR1 structure autorégressive de type 1

BYM Besag-York-Mollié

CAR Modèle autorégressif conditionnel

CMM Communauté métropolitaine de Montréal

CSR *Complete Spatial Randomness*

DIC *Deviance Information Criterion*

DJMA Débit journalier moyen annuel

EPDS Équation partielle différentielle stochastique

FHWA Federal Highway Administration

GRC Gendarmerie royale du Canada

HEC Hautes études commerciales

ICAR Modèle autorégressif conditionnel intrinsèque

iid indépendant et identiquement distribué

KPV kilomètre parcourue par véhicule

LDE Laboratoire de données sur les entreprises

LGCP *Log-Gaussian Cox Process*

LISA *Local Indicators of Spatial Association*

MAUP *Modifiable Areal Unit Problem*

MIDP *Misaligned Data Problem*

MPV Miles parcourus par véhicule

MPS Motif ponctuel spatial

MRC Municipalité régionale de comté

MRP Modèle autorégressif conditionnel intrinsèque

MSc Maîtrise

MTMD Ministère des Transports et de la Mobilité durable

OMS Organisation Mondiale de la Santé

OSM OpenStreetMap

PPH Processus de Poisson homogène

PPNH Processus de Poisson non-homogène

PPS Processus ponctuel spatial

RMR Régions métropolitaines de recensement

RSM *Road Safety Monitor*

SAAQ Société de l'assurance automobile du Québec

SIR *Standard Incidence Ratio*

SPVM Service de Police de la ville de Montréal

SR Secteur de recensement

SQ Sûreté du Québec

THC Delta-9-tétrahydrocannabinol

TIRF *Traffic Injury Research Foundation*

VKT véhicules-kilomètres parcourus

Remerciements

Je tiens à exprimer ma profonde gratitude à mes professeurs, Aurélie Labbe et Nicolas Saunier, pour leur précieuse aide, leur disponibilité et la grande autonomie qu'ils m'ont accordée tout au long de ce projet.

Je remercie également Finn Lindgren pour son aide concise et ses réponses rapides concernant la modélisation ponctuelle, qui ont grandement facilité mes travaux.

Un grand merci à toutes les personnes qui m'ont soutenu face aux défis administratifs, en particulier Sihem Taboubi, directrice de la M. Sc. au HEC, et Guillaume Paré, directeur du Bureau de l'éthique et de l'intégrité en recherche à Polytechnique.

J'aimerais aussi remercier l'ensemble de mes professeurs du département qui m'ont énormément appris au cours de ces dernières années.

Plus personnellement, j'aimerais particulièrement remercier mes parents, Denis et Nancy, pour leur soutien constant au travers de ce long, et parfois éreintant, travail. J'aimerais également remercier mes frère et sœur, Françoise et Bernard, de parfois avoir pris des nouvelles de l'avancement. Enfin, je tiens à exprimer toute ma reconnaissance à ma blonde, Sabrina, qui a été là tout au long de cette aventure et m'a rassuré à de nombreuses reprises.

Introduction

Le 12 mars 2020, le premier ministre du Québec François Legault a tenu une conférence de presse pour annoncer la fermeture des écoles, des cégeps et universités de la province. Le Québec est alors devenu la première province canadienne à déclarer l'état d'urgence sanitaire. Cette annonce a marqué le début de changements durables dans la vie des Québécois, à un point tel qu'aujourd'hui encore, en 2024, certains aspects de la vie ne sont pas revenus à ce qu'ils étaient avant la pandémie, comme le télétravail qui remplace encore partiellement le travail au bureau. Il est d'ailleurs probable que certains aspects de la vie ne reviendront jamais à ce qu'ils étaient avant la pandémie. Les répercussions se sont fait sentir dans de nombreux domaines, notamment la santé, mais aussi l'économie, le climat social, de l'éducation, le transport, etc. Dans ce mémoire, nous aborderons les impacts de la COVID-19 sous le prisme du transport. Les mesures de confinements, restrictions de voyage entre différentes régions, ainsi que les changements de comportements des différents usagers de la route ont assurément modifié la dynamique des déplacements.

En analysant les statistiques disponibles, on constate effectivement des baisses de 20,3% et 33,9% des nombres de blessés graves et légers en 2020 par rapport à la moyenne de 2015 à 2019 (Société de l'assurance automobile du QUÉBEC, 2021). Cependant, le nombre de décès n'a diminué que de 2,2%. Ces premiers constats indiquent des routes avec moins d'accidents légers mais plus d'accidents mortels que ce qui était attendu. En 2021 et 2022, les bilans de la SAAQ indiquent des tendances similaires à celles de 2020, avec des baisses 11% et 4,2% des blessés graves et de 18,5% et 11% pour les blessés légers (Société de l'assurance automobile du QUÉBEC, 2022) (Société de l'assurance au-

tomobile du QUÉBEC, 2023), mais une augmentation de 0,7% et 13,2% des décès.

L'objectif principal de ce mémoire sera de caractériser les impacts de la pandémie de COVID-19 sur les comportements liés à la sécurité routière et sur l'exposition avant et après l'arrivée de la pandémie afin d'éventuellement identifier des interventions pouvant améliorer la sécurité routière Pour ce faire, nous analyserons les accidents entre 2012 et 2022 à l'aide de techniques spatio-temporelles. En raison de contraintes de confidentialité, l'analyse se limitera à la localisation des accidents par municipalité et MRC, sauf pour Montréal, où nous disposons de données plus précises sur les emplacements des accidents. Nous utiliserons des techniques de modélisations surfaciques et zonales utilisant les voisinages entre municipalités et MRC pour modéliser les nombre d'accidents par régions. À Montréal, les données permettront une analyse plus fine et ponctuelle.

Ce mémoire se séparera en quatre grandes sections. La première section propose une revue de littérature de la sécurité routière, alors que la seconde aborde l'analyse et la modélisation de données spatiales. Les données spatiales que nous avons sont ponctuelles et surfaciques. Nous ne nous intéresserons donc que très peu aux données géostatistiques. La troisième section se concentrera sur le prétraitement et l'exploration des données des accidents, mais aussi de sources externes comme *OpenStreetMap* ou les recensements canadiens (extraites à l'aide de la bibliothèque *cancensus*). Enfin, la dernière section présentera les résultats de nos analyses, cherchant à répondre à trois grandes questions :

1. Existe-t-il un effet COVID-19 significatif sur les accidents de la route au Québec ?
2. Quels facteurs expliquent l'augmentation significative des accidents mortels après 2020 ?
3. La localisation des accidents sur l'île de Montréal a-t-elle changée avec la pandémie ?

À travers ce projet, nous espérons mettre en lumière des enseignements précieux qui pourraient aider à guider les décideurs dans d'éventuels politiques de transports sécuritaires et durables. Finalement, ce mémoire vise bien humblement à contribuer à la compréhension des transformations en cours dans le secteur du transport et à ouvrir la voie

à des réflexions sur les futures orientations à adopter dans un monde post-COVID. Il convient de noter que ce mémoire s'inscrit dans le cadre d'un projet plus large sur la sécurité routière au Québec après la pandémie.

Chapitre 1

Revue de la littérature - Sécurité routière

L'objectif de cette revue de littérature sur la sécurité routière est d'introduire les principaux concepts liés à la sécurité sur les routes, ainsi que les facteurs influençant les accidents de la route, tels que les comportements des conducteurs. En particulier, ce chapitre présente un cadre théorique de la sécurité routière, mais se concentre également sur les changements observés sur les routes depuis le début de la pandémie de la COVID-19. L'émergence de la pandémie a en effet conduit à des restrictions de mobilité, à une réduction du trafic routier et à une évolution des comportements des usagers de la route. Cette revue analyse donc les effets de ces changements sur les statistiques d'accidents, tout en mettant en lumière les nouvelles dynamiques de sécurité routière en période de crise sanitaire.

1.1 Sécurité routière et COVID-19

Ce mémoire se concentre principalement sur la modélisation des accidents de la route, en mettant l'accent sur l'analyse statistique et les méthodes de modélisation spatiale et temporelle. Toutefois, une compréhension de base de la sécurité routière est également

essentielle, car elle permet de cerner les enjeux, les facteurs de risque, et les contextes qui influencent les accidents. La sécurité routière est un domaine pluridisciplinaire, intégrant des éléments de politique publique, d'infrastructure, de comportement humain, et de dynamique des véhicules, chacun ayant un impact direct ou indirect sur les risques d'accidents et leur sévérité. Comprendre ces éléments facilite l'interprétation des modèles et des résultats, en reliant les statistiques aux interventions concrètes et aux politiques visant à réduire la fréquence et la gravité des accidents de la route.

1.1.1 Cadre théorique

De manière simple, la sécurité routière se décompose essentiellement en trois composantes : l'exposition, la gravité et le taux d'accident (ELVIK et al., 2009). À la figure 1.1, on illustre les liens entre les composantes, ainsi que certains éléments qui les composent.

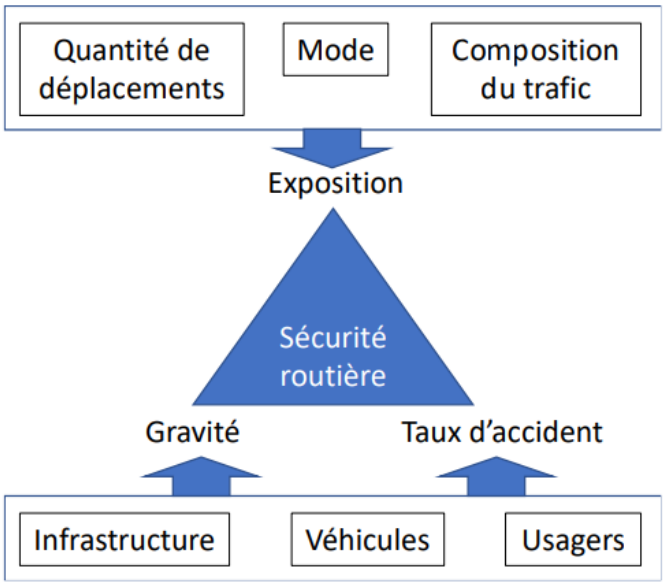

FIGURE 1.1 – Taxonomie des composantes de la sécurité routière. Adaptée de ELVIK et al., 2009 par Nicolas Saunier.

Toujours selon ELVIK et al., 2009, il existerait quatre manières théoriques d'améliorer la sécurité routière :

1. Réduction de l'exposition, donc réduction du nombre et de la quantité de déplacements.
2. Transfert vers des modes de transport plus sécuritaires.
3. Réduction du taux d'accident pour une quantité de déplacements donnée.
4. Réduction de la gravité des accidents en protégeant mieux les individus.

Plongeons maintenant plus en détail dans ces composantes.

Exposition L'exposition représente la quantité d'activités pendant lesquelles un accident peut survenir. Lorsqu'on parle de circulation routière, on utilise souvent la quantité totale des distances parcourues ou quantité de déplacements. Le mode de transport impacte également l'exposition. En effet, un individu à pied, en autobus, en vélo, ou en auto n'encourt pas le même risque d'être impliqué dans un accident. Comme la composition du trafic dépend directement des modes de transport, elle influence elle aussi l'exposition,

Bien qu'il semble simple de calculer la distance parcourue totale, il est en réalité difficile de trouver des données fiables à ce sujet, et ce l'est d'autant plus si on veut inclure d'autres modes de transport que les véhicules motorisés, comme les piétons ou les cyclistes.

L'estimation généralement utilisée repose sur le véhicules-kilomètres parcourus (VKT). Le VKT quotidien peut être estimé à l'aide des débits journaliers moyens annuels (DJMA). Pour une section de route i, on a (TRANSPORTATION et F. H. ADMINISTRATION, 2018) :

$$\text{VKT}_i = \text{DJMA}_i \times \text{longueur}_i. \tag{1.1}$$

Des estimations de DJMA sont généralement accessibles, du moins pour une partie du réseau routier. Pour obtenir le VKT d'une région, il suffit de sommer les VKT de toutes les sections de route dans cette région :

$$\text{VKT}_{\text{région}} = \sum_{i \in I} \text{VKT}_i, \tag{1.2}$$

où I est l'ensemble des sections de route contenues dans la région.

Taux d'accident Le taux d'accident correspond à la probabilité d'être impliqué dans un accident par unité d'exposition. La probabilité de survenance désigne la chance qu'un accident se produise dans un contexte donné. Il s'agit en quelque sorte d'une fréquence. Il est souvent supposé que cette probabilité est proportionnelle à la probabilité d'accident.

La probabilité de survenance d'un accident est expliquée par une multitude de facteurs de risque. Des exemples de ces facteurs de risque sont les infrastructures, les véhicules, les usagers de la route eux-mêmes, ou les dispositifs de contrôle comme les panneaux ou les feux de circulation.

Gravité La gravité représente simplement la sévérité des blessures des individus impliqués dans un accident. En théorie, la gravité est continue. En pratique, elle est souvent catégorisée selon l'intensité de l'accident, telle que mortel, grave ou léger. Puisque ces catégories sont arbitraires, elles sont difficilement comparables entre pays, même s'il existe des échelles internationales pour permettre la comparaison. Un exemple de ces échelles serait l'*Abbreviated Injury Scale* (ADVANCEMENT OF AUTOMOTIVE MEDICINE, s. d.).

1.2 Changements avant et après la COVID-19

Dans le but de comprendre l'impact qu'a eu la COVID-19 sur la sécurité routière, commençons par une observation très générale. Premièrement, les infrastructures et les véhicules n'ont pas beaucoup évolué durant la pandémie. Ce n'est donc pas par ces éléments que la sécurité routière a été affectée. En se basant sur la figure 1.1, il ne reste donc que l'exposition et/ou les comportements des usagers qui auraient pu changer.

En parlant d'exposition, ELVIK et al., 2009 mentionne que, bien qu'une baisse d'exposition entraîne une baisse du nombre d'accidents, il ne s'agit pas d'une relation linéaire. Un exemple, directement tiré du livre, est celui d'une situation avec 500 piétons et 5000 véhicules motorisés. En doublant le volume de trafic, soit en passant à 1000 piétons et 10 000 véhicules motorisés, le nombre d'accidents impliquant des piétons augmente d'un facteur de 2,33 (soit plus de deux fois). Remarquons aussi que, malgré cette augmenta-

tion du nombre d'accidents impliquant des piétons, la probabilité pour un piéton d'être impliqué dans un accident diminue. En effet, en passant de 100 à 1000 piétons, le nombre d'accidents par piéton diminue de 50%. Ce phénomène existe aussi avec les véhicules motorisés et les cyclistes.

À la fois pour les piétons, les véhicules motorisés et les cyclistes, un haut débit confère une meilleure protection individuelle malgré une hausse globale des accidents. Ce phénomène est connu sous le nom de *safety in numbers*. Des hypothèses pour expliquer cela mentionnent que dans un grand volume de trafic, la vitesse diminue et les usagers de la route, supposant qu'ils ne veulent pas être impliqués dans un accident, sont plus attentifs à leur environnement. À l'opposé, on pourrait s'attendre à ce qu'un plus faible achalandage augmente les comportements risqués, ce qui concorderait avec un raisonnement avancé dans VANLAAR et al., 2021.

1.2.1 Infractions

On s'attendrait à ce qu'une hausse des comportements risqués soit associée à une augmentation des infractions, et donc des contraventions données par la police. C'est justement ce que révèlent plusieurs bilans de police, articles de journaux et articles scientifiques qui décrivent les comportements des usagers de la route pendant la pandémie de la COVID-19. Nous résumons cela ci-dessous, en segmentant par grandes régions.

Reste du Canada (Hors-Québec) Au tout début de la pandémie, on remarquait une augmentation des comportements risqués un peu partout à travers le pays. Des données de la ville d'Edmonton montraient une augmentation de 30% du nombre d'excès de vitesse de plus de 20km/h, malgré une baisse de 30% du nombre de véhicules (PHIL, 2020). La police de Toronto a, quant à elle, observé une augmentation de contraventions pour excès de vitesse de 35%, et de presque 200% de contraventions pour *stunt driving*, qu'on peut traduire par « conduite de cascadeurs » (TORONTO, 2020). La police de Saanich en Colombie-Britannique demandait à ses citoyens de ralentir puisqu'elle est passée de deux véhicules mis en fourrière pour vitesse excessive à 16 dans les 30 jours précédents et

suivants le début de la pandémie (ADAM, 2020). Remarquons que même si cette dernière statistique est rapportée dans quelques études, dont LYON, VANLAAR et ROBERTSON, 2024 et VANLAAR et al., 2021, il y a très peu d'observations. Malgré tout, notons que le nombre de véhicules mis en fourrière au Nouveau-Brunswick a lui aussi augmenté de 17% entre 2019 et 2020 (RCMP, 2022) et que ce nombre continue d'augmenter en 2022 (RCMP, 2023).

Regardons maintenant les comportements risqués depuis le déclenchement de la pandémie, mais sur du moyen à long terme. La Gendarmerie royale de Canada (GRC) du Nouveau-Brunswick a remis 16,9% plus de contraventions pour excès de vitesse en 2020 par rapport à 2019. Ce nombre a ensuite diminué de 22,7% en 2021 par rapport à 2020. Par contre, il y a aussi certains endroits qui n'ont pas vu de hausses de contraventions en excès de vitesse, comme Saskatoon qui a vu sa police en distribuer 46% moins en 2020 par rapport à 2019. La section de la police de Saskatoon affectée au trafic avait cependant été redéployée pendant près de 3 mois (THIA, 2021). Encore au Nouveau-Brunswick, les distractions au volant rapportées par la GRC ont d'abord diminué de 18,4% de 2019 à 2020, puis sont restées relativement stables en 2021, et ont ensuite augmenté de 77,8% en 2022. Pour ce qui est maintenant des infractions relatives à la conduite avec facultés affaiblies, la région d'Ottawa a vu une augmentation de 47 à 60 automobilistes arrêtés pour conduite avec facultés affaiblies entre les mois de novembre 2019 et novembre 2020 (JOSH, 2021) sur une base d'environ 650 000 conducteurs à Ottawa (PROVINCE DE L'ONTARIO, 2020). Les données de la GRC du Nouveau-Brunswick ne corroborent pas ceci, on y retrouve des baisses d'infractions relatives à la conduite avec facultés affaiblies depuis 2019.

Québec La Sûreté du Québec (SQ) ne partage que les données relatives aux accidents mortels sur son territoire dans son bilan. Ce qui est surtout remarquable en examinant les bilans de 2018 à 2022, c'est l'augmentation du nombre d'accidents mortels ayant pour cause première probable la distraction, passant de 10% des victimes décédées avant 2020 à 15% après. De plus, l'absence de ceinture de sécurité chez les victimes décédées augmente elle aussi, passant de 10–15% avant 2020, à 20–30% après 2020 (Sûreté du

QUÉBEC, 2021). Pour ce qui est des accidents mortels causés par la conduite imprudente ou vitesse excessive et ceux causés par des facultés affaiblies, la SQ ne rapporte pas de grandes différences de 2018 à 2022.

Le Service de police de la Ville de Montréal (SPVM) rapporte quant à lui le décompte de toutes les infractions liées à la sécurité routière (MONTRÉAL, 2023). Les infractions pour conduite dangereuse ont augmenté de 124% de 2019 à 2022. Les infractions pour conduite avec facultés affaiblies étaient en 2022 de 68 % inférieures à celles de 2019. Toutefois, ce nombre avait d'abord chuté drastiquement de 2019 à 2020, puis a commencé à augmenter depuis.

États-Unis Les données de la *National Highway Traffic Safety Administration* (NHTSA) suggèrent que le nombre d'éjections a augmenté en 2020 et 2021 par rapport à 2019 (N. H. T. S. ADMINISTRATION, 2021). Les éjections peuvent être considérées comme un proxy du port de la ceinture de sécurité. Ce sont les hommes de milieux ruraux chez qui la hausse du nombre d'éjections est la plus marquée. C'est d'ailleurs en milieu rural que les vitesses les plus élevées sont observées sur les autoroutes.

La NHTSA remarque une diminution globale du nombre de collisions, mais signale que l'augmentation des décès chez les piétons requiert une attention particulière. De plus, elle avance que l'augmentation des ventes de cannabis et d'alcool pourrait indiquer un changement social qui pourrait avoir des implications sur le trafic.

TELECOMMUNICATIONS, 2021 rapporte une augmentation de la manipulation du téléphone par kilomètres parcourus en mars 2021 par rapport à janvier 2020 aux États-Unis. Depuis 2020, il y a eu une augmentation de 23% du temps d'utilisation d'un téléphone par heure de conduite. Cette moyenne correspond maintenant à 2m12s d'interaction par heure (TELECOMMUNICATIONS, 2023). Un conducteur impliqué dans un accident a deux fois plus de chance d'avoir utilisé son téléphone dans la minute précédant son accident. En effet, 34% des conducteurs impliqués dans un accident ont interagi avec leur téléphone dans la minute précédant l'accident.

En Alabama, il a été observé qu'une plus grande proportion des collisions était asso-

ciée à des comportements risqués, tels que la conduite avec facultés affaiblies, l'absence de ceinture de sécurité, ou la conduite distraite, pendant les périodes de confinement. Les excès de vitesse y étaient cependant plus fréquents avant le confinement (ADANU et al., 2022).

Une étude très intéressante (DONG, XIE et YANG, 2022) met de l'avant que l'augmentation de deux variables sous-jacentes, l'agressivité et l'inattention, explique l'augmentation de probabilité d'accidents plus graves pendant la pandémie en Virginie. Les modèles présentés montraient que l'agressivité et l'inattention ont significativement augmenté après le déclenchement de la pandémie. L'agressivité est identifiée via les facteurs suivants notés dans les rapports d'accident : excès de vitesse, conduite avec facultés affaiblies et dépassement inapproprié. L'inattention, quant à elle, est identifiée par l'absence de ceinture de sécurité, la conduite distraite et l'omission de signaler.

Ailleurs dans le monde Une étude basée sur des données de Grèce et d'Arabie Saoudite (KATRAKAZAS et al., 2020) a montré qu'au cours des deux premiers mois de la pandémie, il y a eu des augmentations de vitesse, du nombre de freinages et d'accélérations brusques ainsi que de manipulations de téléphone.

Conclusion sur les infractions Finalement, malgré des indications qui ne vont pas toujours dans le même sens, les excès de vitesse, les distractions, l'absence de ceinture de sécurité et la conduite avec facultés affaiblies semblent avoir augmenté, du moins à certains moments et dans certains endroits. Cette tendance pourrait être expliquée par des changements dans les comportements des usagers de la route en réponse aux conditions inédites créées par la pandémie, notamment la réduction du trafic, l'isolement social et l'augmentation des comportements risqués en raison de l'absence de surveillance directe. Toutefois, il est important de noter que ces augmentations ne sont pas uniformes et varient selon les régions et les périodes, ce qui suggère que plusieurs facteurs locaux et contextuels ont influencé ces comportements. Ainsi, bien que des conclusions générales puissent être tirées, une analyse plus fine et contextuelle est nécessaire pour mieux com-

prendre l'impact de la pandémie sur la sécurité routière à long terme.

Tout ceci mène à une autre question. Les données sur les infractions émises pendant la pandémie sont-elles fiables ? En effet, des effectifs policiers habituellement affectés à la circulation ont pu être redéployés ailleurs en début de pandémie, comme le mentionnait THIA, 2021. LUM, MAUPIN et STOLTZ, 2020 montre qu'une majorité de services de police ont réduit leur présence proactive sur les routes et ont donc réduit le nombre de contrôles routiers. Cynthia LUM, Carl MAUPIN et Megan STOLTZ, 2022 montre qu'au moins quelques services de police ont rapidement réajusté leur présence proactive sur les routes, en raison de l'augmentation de conduites dangereuses en mai 2020. Il est malheureusement difficile d'obtenir des données québécoises, ou même canadiennes concernant l'affectation des policiers.

1.2.2 Exposition

Suite à la déclaration de la pandémie de COVID-19 par l'Organisation mondiale de la santé (OMS) en mars 2020, des mesures ont été mises en place pour réduire la transmission du virus, et plusieurs de ces mesures cherchaient à réduire la mobilité des individus. La prochaine section traitera de l'exposition, et en particulier, de cette chute d'exposition au Canada et aux États-Unis.

Canada L'Association des industries de l'automobile du Canada (AIA) a publié des données concernant l'achalandage sur les routes, provenant de INRIX Trip Trends (A. d. i. d. l. d. CANADA, 2023) et StreetLight (CANADA ET STREETLIGHT, 2022). Le tableau 1.1 présente les mesures de ces deux sources.

INRIX Trip Trends utilise un indice de kilomètres parcourues par véhicule (KPV) dont le niveau de référence est fixé au début de l'année 2020, supposé représenter une période typique. Le KPV est une mesure moyenne, qui représente le nombre moyen de kilomètres parcourus par chaque véhicule individuel sur une période donnée. Il est calculé en divisant le VKT par le nombre total de véhicules ayant circulé pendant cette période. Les mesures récoltées par StreetLight sont des estimations des véhicules-kilomètres parcourus (VKT)

AIA				
INRIX Trip Trends	**Moyenne des indices de kilomètres parcourus par véhicules (KPV)**			
	Avril 2019 - Mars 2020	Avril 2020 - Mars 2021	Avril 2021 - Mars 2022	Avril 2022 - Mars 2023
Indice KPV	-	0.71	0.64	0.86
Différence (%) avec l'année précédente	-	-	-9.2%	33.9%
StreetLight	**Véhicules kilomètres parcourues (VKT) en milliards**			
	Avril 2019 - Mars 2020	Avril 2020 - Mars 2021	Avril 2021 - Mars 2022	Avril 2022 - Mars 2023
VKT (en milliards)	457	396	457	-
Différence (%) avec l'année précédente	-	-13.3%	15.3%	-

TABLEAU 1.1 – Mesures sur les KPV suite au déclenchement de la pandémie. Source AIA

sur une période donnée, où les VKT sont à peu près égaux à la longueur moyenne d'un trajet multipliée par le nombre total de trajets, donc simplement, la somme des longueurs de tous les trajets sur l'année. Les mesures présentées dans le tableau 1.1 vont d'avril à mars de l'année suivante. Cette période est choisie puisque la pandémie s'est déclenchée en mars 2020. Les données vont de 2019 à 2021 pour StreetLight et de 2020 à 2022 pour INRIX. Nous avons sommé les mesures pour INRIX et fait une moyenne par période pour StreetLight, afin de conserver l'échelle de l'indice. Le tableau montre aussi la variation en pourcentage d'une période à l'autre.

Les données de StreetLight confirment une baisse de la mobilité après le déclenchement de la pandémie. Cependant, les deux sources de données ne concordent pas concernant le moment de la hausse après 2020. En effet, StreetLight montre que les VKT sont revenus à leur niveau pré-pandémique en 2021, alors que INRIX Trip Trends indique que les KPV ont continué de diminuer en 2021 avant de remonter en 2022.

Des données provenant de S. CANADA, 2022 montrant les ventes brutes d'essence au Canada sont présentées au tableau 1.2. L'essence brute vendue peut être utilisé comme proxy de la quantité de déplacements.

On remarque ici aussi une baisse similaire à celle des données de Street light, mais pas à celle d'INRIX, comprise entre 13% et 14% de 2019 à 2020, année du déclenchement de la pandémie. Le retour vers la normalité se fait ici plus progressivement, avec des hausses d'environ 5% en 2021 et 2022. Ces hausses sont significatives, comme on peut le voir en regardant la vente brute d'essence depuis 2000, présentée à la figure 1.2. On voit très

Statistique Canada				
Année	2019	2020	2021	2022
Vente brutes d'essence (millions de litres)	44806	38598	40191	42459
Différence (%) avec l'année précédente	0.0%	-13.9%	4.1%	5.6%

TABLEAU 1.2 – Mesures provenant de Statistique Canada montrant les ventes brutes d'essence au Canada en million de litres.

bien que la chute en 2020 est exceptionnelle, mais que le retour vers la normale est bien entamé et qu'il est plus abrupt que la hausse progressive que l'on observait depuis 2000.

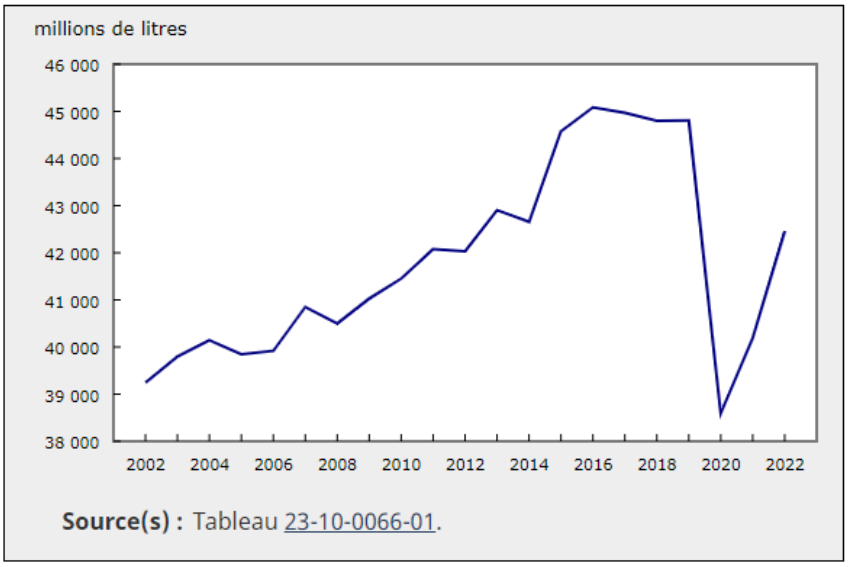

FIGURE 1.2 – Graphique des ventes brutes d'essence au Canada par année, tirée directement de S. CANADA, 2022.

Les mesures de KPV de StreetLight, de VKT d'INRIX Trip Trends et celle d'essence vendue de Statistique Canada montrent toutes qu'il y a eu un déclin abrupt en 2020 et que le retour à la normale est amorcé. Cependant, le moment et l'évolution de ce retour à la normale sont différents selon la source. On a donc ici un exemple montrant qu'il est difficile d'estimer l'exposition de manière satisfaisante. Malgré tout, les VKT de StreetLight et les ventes brutes d'essences de Statistique Canada sont plus semblables.

Il est important de noter que les différentes régions ne suivent pas nécessairement les mêmes tendances au niveau de la quantité de déplacements. Cela est démontré par le tableau 1.3.

Période	Canada	Colombie-Britannique	Prairies	Ontario	Québec
T1 2021	-29.4%	-23.7%	-18.9%	-35.4%	-19.4%
T2 2021	1.9%	-6.0%	-1.3%	-0.6%	20.9%
T3 2021	-28.8%	-36.5%	-27.5%	-28.3%	-22.6%
T4 2021	-16.6%	-27.4%	-21.1%	-15.7%	-6.7%
T1 2022	19.3%	12.0%	-0.3%	23.8%	20.8%
T2 2022	54.9%	58.9%	27.0%	74.9%	41.4%
T3 2022	49.6%	65.8%	25.1%	47.6%	36.9%
T4 2022	36.2%	45.8%	20.7%	32.9%	30.9%
T1 2023	0.6%	2.9%	-4.8%	-1.2%	-2.0%

Variation en % calculée par rapport au même trimestre de l'année précédente

TABLEAU 1.3 – Variation nationale et par région de l'indice KPV par rapport au trimestre de l'année précédente. Tableau tiré directement de A. d. i. d. l. d. CANADA, 2023.

Les différentes régions canadiennes n'ont en général pas les mêmes variations de KPV par rapport au trimestre de l'année précédente. Le Québec se démarque particulièrement au second trimestre de 2021.

Québec En se concentrant particulièrement sur le Québec, des données sur la mobilité vers les lieux de travail provenant du Laboratoire de données sur les entreprises (LDE) indiquent une baisse de 5,2% en février 2024 par rapport à janvier 2020 (ENTREPRISES, 2024). Cette baisse de la mobilité vers les lieux de travail est présentée pour différentes grandes villes du Québec au tableau 1.4. On y fait aussi la distinction entre les centre-villes et les régions métropolitaines de recensement (RMR).

On remarque que, dans tous les cas, la mobilité vers les lieux de travail au centre-ville est plus éloignée du niveau prépandémique que si l'on inclut toute la RMR. De plus, les plus grandes villes comme Montréal, Québec et Gatineau, et les banlieues de Montréal comme Laval et Longueil, avaient encore, en février 2024, une plus faible mobilité vers les lieux de travail que les villes de taille moyenne.

Même s'il était question de provinces au tableau 1.3, et qu'il est question de villes québécoises dans 1.4, les deux conduisent à une conclusion semblable : différentes régions n'ont pas les mêmes variations d'exposition. Bien entendu, les mesures des différentes

Ville	Variation (%) de la mobilité au travail	
Période	Février 2024 par rapport à janvier 2020	
Région	Centre-ville	RMR
Montréal	$-31,0$	$-9,1$
Québec	$-21,9$	$-7,7$
Saguenay	$-6,7$	$3,0$
Sherbrooke	$-6,6$	$-1,2$
Trois-Rivières	$-5,3$	$5,2$
Gatineau	$-22,9$	–
Laval	$-58,2$	–
Longueuil	$-27,0$	–

TABLEAU 1.4 – Variation de la mobilité vers le lieu de travail de février 2024 par rapport à janvier 2020 selon la ville. Les données marquées d'un « – » ne sont pas disponibles. Ces données proviennent de ENTREPRISES, 2024.

sources ne sont pas toutes comparables.

CHAN et al., 2020 indiquait déjà en novembre 2020 que les régions plus peu enclines à prendre des risques ont ajusté leurs comportements suite au déclenchement de la pandémie, et ce, avant même que les gouvernements imposent des confinements. Les régions moins sensibles au risque étaient quant à elles plus susceptibles de maintenir leur mobilité. Le risque dans cette étude était un risque général, n'ayant pas directement rapport au risque routier. Il a été mesuré à l'aide de sondages.

Bref, on peut conclure avec certitude qu'il y a eu une baisse de la quantité totale de déplacements au Canada et au Québec au déclenchement de la pandémie. Différentes régions n'ont pas répondu au risque de la même manière, donc la baisse de déplacements n'était pas la même partout. Finalement, le retour à la normale est bien entamé en 2022 et 2023, mais il est très difficile de déterminer où il en est, surtout que celui-ci est sûrement différent d'une région à l'autre. Cela soulève une question importante : qu'entend-on par « normale » ? Devrait-t-on s'attendre à un retour au niveau prépandémique ? Bien que cette question soit intéressante, nous nous concentrerons ici sur la rupture provoquée par la pandémie.

États-Unis Les données sur la mobilité aux États-Unis sont aussi très intéressantes. On s'attendrait à ce que ces données ressemblent à celles du Canada. La *Federal Highway Administration* (FHWA) publie d'ailleurs des statistiques sur les miles parcourus par véhicule (MPV), présentées à la figure 1.3.

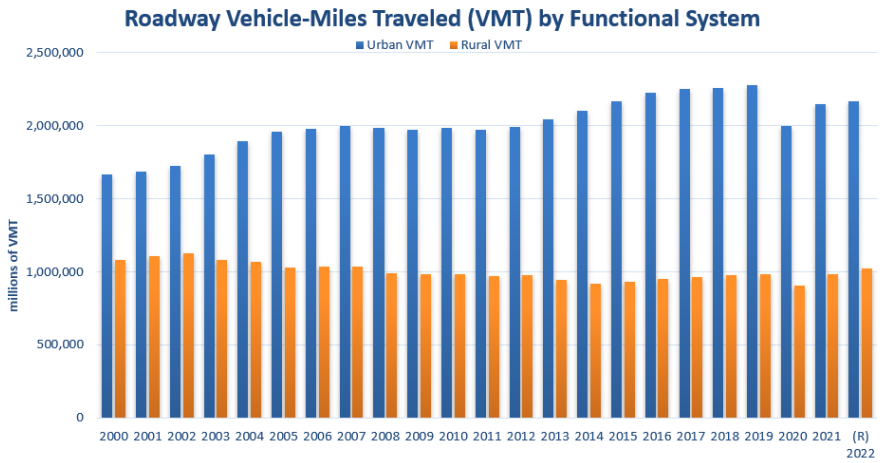

FIGURE 1.3 – Miles parcourus par véhicule (MPV) de 2000 à 2022, par région rurale ou urbaine aux États-Unis. Graphique tiré directement de TRANSPORTATION STATISTICS, 2022.

Que ce soit en région urbaine ou rurale, on peut voir la baisse en 2020, suivie d'un retour en 2021 et 2022 vers la normale. En région rurale, on retrouve des niveaux prépandémiques en 2022, ce qui n'est pas le cas en région urbaine. Ceci pourrait être expliqué par des changements dans la mobilité vers les lieux de travail, qui, au Canada, sont plus faibles dans les centres-villes que dans d'autres régions (ABDOU, 2023). Le tableau 1.5 montre les mesures de MPV des années 2019 à 2022, ainsi que la variation par rapport à l'année précédente.

Notons qu'aux États-Unis aussi, différentes régions (représentées par les classes urbaines ou rurales) réagissent différemment en termes de mobilité. Finalement, on voit la même tendance générale qu'au Canada, soit la chute au déclenchement de la pandémie, et un retour vers la normale depuis.

FHWA					
Année		2019	2020	2021	2022
Urbain	VMT (en millions)	2,277,919	2,000,095	2,148,130	2,170,481
	Différence (%) avec l'année précédente	0,7%	-12.2%	7.4%	1.0%
Rural	VMT (en millions)	983,853	903,527	984,281	1,025,711
	Différence (%) avec l'année précédente	0,5%	-8.2%	8.9%	4.2%

TABLEAU 1.5 – [MPV de 2019 à 2022 aux États-Unis, en région rurale et urbaine.] MPV de 2019 à 2022 aux États-Unis, en région rurale et urbaine. Données provenant de TRANSPORTATION STATISTICS, 2022.

1.2.3 Étude canadienne à propos de comportements de la route autodéclarés

Une étude de LYON, VANLAAR et ROBERTSON, 2024 a analysé les comportements sur la route des Canadiens avant et pendant la pandémie de COVID-19. Les données ont été récoltées dans le cadre du projet *Road Safety Monitor* (RSM) par le *Traffic Injury Research Foundation* (TIRF). Le RSM est un sondage public annuel qui questionne les canadiens sur plusieurs problématiques de sécurité routière. Les données du RSM sont fortement corrélées aux données relatives aux accidents mortels (FOUNDATION, 2023). Le sondage utilisé dans cette étude questionnait les répondants à propos de leurs comportements sur la route entre mars 2021 et mars 2022. Au total, il y a eu 2978 réponses, dont 1768 conducteurs et 1210 non-conducteurs. Bien sûr, les questions sur la conduite n'étaient demandées qu'aux conducteurs, soit les personnes ayant conduit une auto au cours des 30 derniers jours. Les éléments qui ressortent particulièrement de l'étude sont une hausse des comportements risqués et un changement dans les modes de transport, avec, entre autres, une hausse de la marche comme mode de transport.

Comportements risqués

Tout d'abord, l'étude révèle que 10,1% des conducteurs ont trouvé qu'il était plus difficile de se concentrer pendant la pandémie par rapport à avant. Il serait logique de penser que ceci se traduit par plus de conducteurs distraits. Une conduite considérée comme plus

distraite est une conduite où le conducteur s'adonne à des tâches non-liées à la conduite, comme faire des appels, texter, surveiller des enfants, etc. Paradoxalement, 2,8% des conducteurs ont rapporté conduire plus souvent en étant distraits et lorsqu'on a demandé s'ils observaient d'autres conducteurs distraits, 28,7% ont répondu qu'ils en observaient plus souvent qu'avant. On remarque le même phénomène paradoxal avec les excès de vitesse. En effet, en demandant aux conducteurs s'ils avaient fait des excès de vitesse de plus de 20km/h au-dessus de la limite permise, seulement 2,9% ont répondu qu'il le faisait plus souvent, et pourtant, en demandant s'ils avaient observé d'autres conducteurs faisant de tels excès de vitesse, 30,6% ont répondu qu'ils en observaient plus souvent. De plus, en demandant si les conducteurs avaient été impliqués dans des situations pouvant mener à un accident, 11,9% ont répondu que c'était arrivé plus souvent, alors que 28,7% ont observé d'autres conducteurs dans des situations pouvant mener à un accident plus souvent. On peut peut-être attribuer ces écarts entre comportements risqués de soi et observation de ceux des autres à la conception erronée répandue qu'on est un conducteur au-dessus de la moyenne. D'autres résultats de l'étude par rapport aux comportements risqués sont présentés au tableau 1.6.

Étonnamment, contrairement à ce qui est rapporté dans les données policières, l'utilisation de la ceinture de sécurité rapportée aurait augmenté, avec une hausse de 11,0% de conducteurs la portant plus souvent. Les résultats de l'étude montrent aussi qu'une proportion non négligeable de personnes affirme avoir consommé plus d'alcool ou de cannabis contenant du THC pendant la pandémie. Ces hausses sont respectivement de 14,9 et 12,2 points de pourcentages. On remarque d'ailleurs que près de 10% des répondants conduisent plus souvent dans les deux heures après avoir consommé alcool et/ou drogues pendant la pandémie.

Il est important de noter qu'une majorité des conducteurs n'a rapporté aucun changement à leurs comportements sur la route. Ceci est intéressant puisque ces personnes n'ont pas ajusté leur prise de risque, malgré les appels répétés à rester plus vigilants pour ne pas surcharger le système de la santé. Cependant, le segment de population ayant augmenté sa fréquence de comportements risqués est plus inquiétant. L'une des hypothèses

Behavior	More Often	No Change	Less Often	Never Done So
Drive within 2 hrs of drinking alcohol	8.7 (7.1–10.6)	47.3 (44.3–50.3)	6.0 (4.7–7.6)	38.0 (35.2–41.0)
Drive when probably over the legal alcohol limit	6.6 (5.2–8.3)	38.8 (35.9–41.8)	4.2 (3.2–5.6)	50.4 (47.4–53.4)
Drive within 2 hrs of using drugs	5.2 (4.1–6.6)	34.8 (32.4–37.3)	4.0 (3.1–5.2)	56.0 (53.4–58.6)
Drive within 2 hrs of using both alcohol and drugs	9.8 (6.7–14.1)	33.4 (28.2–39.0)	12.1 (8.7–16.5)	44.8 (39.1–50.7)
Exceed posted speed limit by 20km/hr	2.9 (2.1–3.9)	80.6 (78.4–82.6)	16.6 (14.7–18.6)	n/a[1]
See other drivers speeding excessively	30.6 (28.8–32.5)	62.4 (60.5–64.4)	6.9 (5.9–8.0)	n/a[1]
Drive while distracted	2.8 (2.1–3.9)	82.2 (80.0–84.1)	15.0 (13.2–17.0)	n/a[1]
See other drivers drive distracted	28.7 (26.9–30.6)	64.9 (63.0–66.8)	6.4 (5.5–7.4)	n/a[1]
Drive while fatigued	3.6 (2.8–4.7)	82.4 (80.3–84.4)	14.0 (12.2–15.9)	n/a[1]
Wearing a seatbelt	11.0 (9.8–12.3)	86.4 (84.9–87.7)	2.6 (2.1–3.4)	n/a[1]

1 Option of 'never done so' not provided.

TABLEAU 1.6 – Comportements risqués autodéclarés pendant la pandémie. Tiré directement de LYON, VANLAAR et ROBERTSON, 2024.

évoquées dans LYON, VANLAAR et ROBERTSON, 2024 est que ces individus percevaient les risques dans la conduite comme faisant partie de leur identité, et ont eu une réponse défensive par rapport aux appels à diminuer le risque en général. Cette conclusion est semblable à celle présente dans AGRAWAL et DUHACHEK, 2010, qui étudiait la réponse aux appels contre la surconsommation d'alcool. On y observait que ces appels étaient moins efficaces s'ils causaient de la honte ou de la culpabilité. D'autres hypothèses soulevées dans LYON, VANLAAR et ROBERTSON, 2024 suggèrent que les individus ayant été plus impactés pendant la COVID-19 par des fermetures d'entreprises, des pertes d'emploi ou un filet social plus restreint sont ceux qui prennent plus de risques.

Mode de transport

Les modes de transport utilisés pendant la pandémie ont significativement changé. Comme mentionné plus tôt, le mode de transport piéton est celui qui a le plus augmenté, avec 29% des individus indiquant qu'ils marchaient davantage. On peut voir ce résultat ainsi que les autres dans le tableau 1.7.

Mode	Less Time	No Change	More Time	Yes to Long-term change	No to Long-term change
Driving Or Riding as a Passenger	21.1 (19.5–22.8)	67.7 (65.8–69.5)	11.2 (10.0–12.5)	58.7 (55.2–62.1)	41.3 (37.9–44.8)
Taxi or Rideshare	19.1 (17.6–20.8)	75.1(73.3–76.8)	5.8 (4.9–6.8)	61.6 (57.5–65.5)	38.4 (34.5–42.5)
Walking	10.9 (9.7–12.2)	60.1 (58.1–62.1)	29.0 (27.2–30.8)	72.6 (69.7–75.4)	27.4 (24.6–30.3)
Public Transit	23.8 (22.1–25.6)	69.7 (67.8–71.5)	6.5 (5.6–7.6)	63.4 (59.8–67.0)	36.6 (33.0–40.2)
Bicycle	9.7 (8.5–11.0)	83.5 (82.0–85.0)	6.8 (5.8–7.8)	63.0 (58.0–67.8)	37.0 (32.2–42.0)

TABLEAU 1.7 – Variations autodéclarées des modes de transport pendant la pandémie. Les valeurs entre parenthèses sont les intervalles de confiance à 95%. Tiré directement de LYON, VANLAAR et ROBERTSON, 2024.

Encore ici, on remarque que la majorité des sondés n'a pas changé ses habitudes. De

plus, à l'exception des piétons, les répondants ont rapportés des baisses plus importantes que des augmentations dans l'utilisation de tous les modes de transport. Parmi ces baisses, la plus marquée concerne l'utilisation du transport en commun, ce qui s'explique par la volonté de réduire les situations impliquant plusieurs personnes dans des espaces confinés. Cependant, parmi ceux qui ont changé leurs habitudes, une majorité indiquait que ces changements allaient être permanents.

Enfin, il convient de noter que le transfert à différents modes était attribuable à des facteurs reliés à la pandémie, plutôt qu'aux raisons habituelles (coût, temps et confort). Les effets à long terme des changements dans les modes de transport survenus pendant la pandémie doivent être pris en compte pour de futures mesures de sécurité routière.

1.3 Synthèse de la revue en sécurité routière

En général, on observe une augmentation des comportements risqués, comme les excès de vitesse et la conduite distraite ou avec facultés affaiblies. Certaines personnes ont d'ailleurs rapporté avoir consommé plus de drogue et/ou alcool pendant la pandémie. Les bilans policiers suggèrent une diminution du port de la ceinture de sécurité, mais selon les sondés canadiens, on devrait plutôt voir une augmentation du port de la ceinture de sécurité. Par ailleurs, le mode de transport ayant le plus augmenté en utilisation par rapport à avant la pandémie est la marche, alors que celui ayant le plus baissé est le transport en commun. Aux États-Unis, cette tendance est d'autant plus préoccupante, avec une hausse significative des décès chez les piétons. Finalement, une tranche de la population déclare avoir pris plus de risques pendant la pandémie qu'avant, et cette tranche pourrait peut-être, à elle seule, expliquer l'augmentation observée des comportements risqués.

Chapitre 2

Revue de la littérature - Analyse spatiale, temporelle et spatio-temporelle

L'objectif de ce second chapitre de revue de littérature est d'introduire les principaux sujets de ce mémoire. Il y sera question de données spatiales, d'analyses spatiales et d'analyses spatio-temporelles. Pour illustrer certains concepts, nous utiliserons des données québécoises et montréalaises sur les accidents. Les données montréalaises fournissent la localisation exacte des accidents, tandis que les données québécoises se limitent à indiquer la municipalité où l'accident a eu lieu.

2.1 Données spatiales

Les données spatiales sont des données contenant de l'information sur une entité, un évènement ou tout autre objet rattaché à de l'information spatiale, pouvant prendre la forme d'un point (coordonnées), mais aussi la forme de ligne (routes) ou bien de région (municipalités). Ce type de donnée est utilisé dans plusieurs domaines tels que l'environnement, la santé publique, l'économie, ou bien la sécurité routière. Les données spatiales peuvent être considérées comme des observations d'un processus stochastique que nous définirons plus tard. Des exemples de données spatiales sont la température à un point spa-

tial, ou bien le nombre d'accidents mortels dans une municipalité donnée. Notons qu'ici, la municipalité pourrait être représentée par les coordonnées de son centroïde, mais plus souvent, on préfère la représenter comme une zone ayant des liens avec ses voisins.

Selon N. CRESSIE, 1993, il existe trois types de données spatiales. Il y a les données géostatistiques, surfaciques et les motifs ponctuels spatiaux (MPS ou *spatial point patterns* en anglais). Dans cette section, nous approfondirons ces trois types, en particulier en examinant leur domaine respectif, car c'est là où on trouve leur principale différence.

2.1.1 Données surfaciques

Les données surfaciques sont définies sur des sous-régions du domaine d'étude. Le domaine d'étude, noté D, correspond à la région complète. Ce domaine est fixe et partitionné en un nombre fini de sous-ensembles. Les données surfaciques sont généralement des agrégations d'une variable d'intérêt par sous-région. Un exemple de données surfaciques est présenté à la figure 2.1. Il s'agit du nombre d'accidents mortels dans des régions administratives du sud du Québec en 2020.

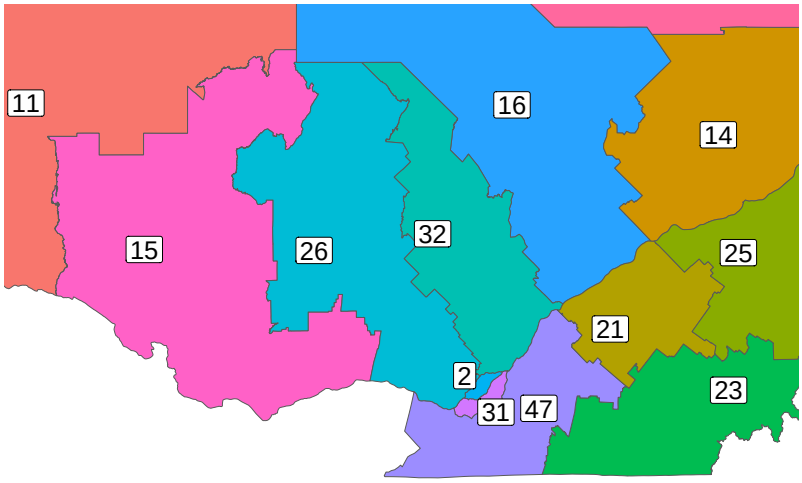

FIGURE 2.1 – Exemple de données surfaciques : nombre d'accidents mortels de régions administratives du sud du Québec en 2020.

2.1.2 Données géostatistiques

Les données géostatistiques sont des données spatiales observées en des points du domaine D. Ce sont des observations à support ponctuel, qui peuvent être prises à n'importe quel endroit du domaine. Un exemple classique de ce type de données serait la température mesurée à plusieurs stations sur un territoire donné. Une station météorologique peut en effet se trouver n'importe où dans une région. La localisation des points est fixe, mais on y observe un attribut Z au point s, noté $Z(s)$ provenant d'un phénomène spatialement continu.

La localisation des accident est une donnée à support ponctuel, mais étant donné que cette localisation est aléatoire et ne provient pas d'un phénomène spatialement continu, les données localisées d'accidents ne peuvent pas être considérées comme des données géostatistiques, il s'agit alors plutôt d'un motif ponctuel spatial. Néanmoins, certaines techniques d'analyse géostatistique peuvent être appliquées à ces données d'accidents, bien que nous ne les approfondirons pas dans ce mémoire.

2.1.3 Motif ponctuel spatial

Un motif ponctuel spatial (MPS) est un ensemble de points généré par un processus ponctuel spatial (PPS). Le PPS est ce qu'on modélise, et le MPS n'est qu'une réalisation du PPS. La localisation des points d'un MPS est aléatoire, alors qu'elle est fixe dans les données géostatistiques. Le domaine D est seulement composé des réalisations. L'attribut observé $Z(s)$ au point s indique alors simplement l'occurrence d'un évènement. On a donc que $Z(s) = 1$, pour tout $s \in D$. Un exemple de données ponctuelles est constitué par les accidents mortels ayant eu lieu à Montréal entre le 1^{e} janvier 2015 et le 31 décembre 2021, présenté à la figure 2.2.

Il est probable qu'en plus de récolter la localisation précise d'un évènement, on récolte aussi certains attributs relatifs à cet évènement. On parle alors de motif ponctuel spatial avec marque. Par exemple, pour un point représentant l'occurrence d'un accident, on pourrait également récolter la date, ou bien le nombre de victimes de l'accident. Les va-

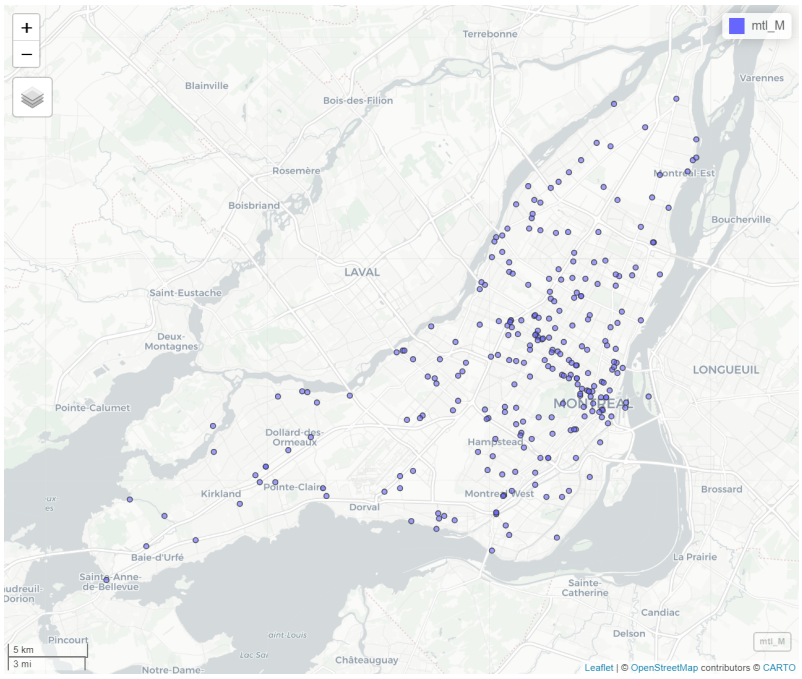

FIGURE 2.2 – Données ponctuelles : accidents mortels à Montréal de 2015 à 2021

leurs associées aux points spatiaux sont appelées des marques. Un MPS avec une marque donnant une indication de temps peut éventuellement être vu comme un processus ponctuel spatio-temporel (BADDELEY, RUBAK et TURNER, 2015). Généralement, lorsqu'on analyse des MPS, on modélise la localisation des points, et non les attributs de ces points. Il est important de différencier les marques des covariables. Les marques sont associées à des points spatiaux et font donc partie de la réponse, tandis que les covariables font, au contraire, partie de l'explication. Des covariables spatiales telles que la densité de population ou l'occupation du sol peuvent être utilisées comme variables explicatives de la localisation des points.

Nous venons de mettre l'accent sur la différence entre covariables et marques. Dans cette recherche, nous présenterons quelques techniques et modèles permettant d'inclure les marques dans la modélisation. Notons également qu'il est possible de modéliser des marques en considérant la localisation des points comme variable explicative. On pourrait par exemple avoir un modèle visant à expliquer la gravité d'un accident en fonction de la localisation de cet accident. On utiliserait alors des techniques d'analyse géostatistique.

2.2 Processus spatiaux

Deux des trois types de données spatiales, soient les données géostatistiques et les MPS, sont à support ponctuel. Nous rappelons qu'avec des données géostatistiques, la localisation est fixe, alors qu'avec des MPS, elle est aléatoire. La différence majeure entre les deux provient du fait que la localisation des points n'est pas d'intérêt dans les données géostatistiques, alors qu'elle est centrale dans les MPS. Dans plusieurs cas, on suppose que les données spatiales à support ponctuel proviennent de processus stochastiques. Commençons d'abord par définir un processus stochastique pour ensuite décrire quelques processus spatiaux importants.

2.2.1 Processus stochastique

D'abord, introduisons un processus stochastique. Un processus stochastique est une collection de variables aléatoires notée $Z(s)$ pour $s \in D \subset \mathbb{R}^d$, où s est l'index d'un point spatial, temporel ou spatio-temporel dans D, un espace à d dimensions. Si la variable $Z(s)$ est continue sur le domaine, alors le processus stochastique est à index continu.

Dans un contexte spatial, on dit d'un processus stochastique qu'il est faiblement stationnaire si ses propriétés statistiques, telles que la moyenne et la variance, demeurent constantes sur le domaine D et que la covariance entre deux points dépend uniquement de la distance qui les sépare.

2.2.2 Processus gaussien

Un processus stochastique est un processus gaussien si n'importe quel ensemble de variables aléatoires du processus suit une distribution normale multivariée. Un processus gaussien $Z(\cdot)$ est défini par des fonctions de moyenne $\mu(\cdot)$ et de covariance $k(\cdot,\cdot)$. Plus précisément, pour un ensemble de n points de l'espace D, notés $s_1, s_2, \ldots, s_n$, nous avons :

$$[Z(s_1), Z(s_2), \cdot, Z(s_n)] \sim \mathscr{MN}(\boldsymbol{\mu}, \boldsymbol{K}), \tag{2.1}$$

tels que $\boldsymbol{\mu} = (\mu(s_1), \mu(s_2), \ldots, \mu(s_n)$ et $\boldsymbol{K} = [k(s_i, s_j)]_{i,j=1,2,\cdot,n} \in \mathbb{R}^{n \times n}$.

Dans cette recherche, nous serons souvent amenés à modéliser des effets spatiaux à l'aide de processus gaussiens. Cette section posera les bases à la compréhension de ce type de processus. Il existe plusieurs fonctions de covariance, nous en présenterons trois. Celles-ci supposent toutes que la covariance est stationnaire, puisque cela facilite la modélisation. Comme la stationnarité d'un processus stochastique, une fonction de covariance est dite stationnaire si elle dépend uniquement de la distance entre deux points $d(s,s')$.

Covariance exponentielle

La fonction de covariance exponentielle décrit une covariance qui décroît exponentiellement avec la distance entre deux points. Elle est définie ainsi :

$$k(s,s') = \sigma^2 \exp(-\kappa||s-s'||), \tag{2.2}$$

où $||s-s'||$ est la distance entre les points s et s'. Généralement, en 2D, on utilise la distance euclidienne. Le paramètre σ^2 représente la variance du processus gaussien, et finalement, κ est un paramètre d'échelle strictement positif qui contrôle à quel point la corrélation diminue avec la distance. Plus précisément, κ est défini comme suit :

$$\kappa = \frac{\sqrt{8\nu}}{\rho}. \tag{2.3}$$

Le paramètre ρ est appelé portée et il représente la distance à laquelle la corrélation n'est plus que de 0,14 (KRAINSKI et al., 2020) et $\nu > 0$ est un paramètre de lissage. Les paramètres sont présentés en détail afin de faire des liens par la suite.

Covariance exponentielle quadratique

La fonction de covariance exponentielle quadratique est très semblable à la fonction de covariance exponentielle, et prend cette forme :

$$k(s,s') = \sigma^2 \exp(-(\kappa||s-s'||)^2). \tag{2.4}$$

Les processus résultants sont très lisses, et c'est ce qui rend cette fonction de covariance irréaliste dans plusieurs processus physiques. Pourtant, la fonction exponentielle quadratique est la fonction de covariance la plus couramment utilisée dans les domaines de *kernel machines*, donc dans des techniques comme les machines à vecteurs de supports (SVM) (RASMUSSEN et WILLIAMS, 2006).

Covariance de Matérn

Les processus trop lisses engendrés par la fonction exponentielle quadratique nous poussent vers la fonction de covariance de Matérn. Il s'agit d'une généralisation des fonctions exponentielle et exponentielle quadratique, mais aussi d'une fonction de covariance très flexible qui se manifeste dans plusieurs domaines (EMILIO et al., 2024). Cette fonction de covariance est généralement utilisée dans l'analyse spatiale (WIKLE, ZAMMIT-MANGION et CRESSIE, 2019). Elle est définie comme suit :

$$k(s,s') = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||s-s'||)^{\nu}K_{\nu}(\kappa||s-s'||), \tag{2.5}$$

où s, s', σ^2, ν et κ sont les mêmes que ceux définis à l'équation 2.2, $\Gamma(\cdot)$ est la fonction gamma et $K_{\nu}(\cdot)$ est la fonction Bessel modifiée du second ordre.

Le paramètre ν contrôle la différentiabilité de la fonction $k(s,s')$. En effet, plus la valeur de ν est élevée, plus la fonction de covariance devient lisse et différentiable. En fixant ν à $\frac{1}{2}$, on retrouve la fonction de covariance exponentielle, qui n'est pas différentiable partout. Lorsque $\nu \to \infty$, on retrouve la fonction de covariance exponentielle quadratique, qui est infiniment différentiable. À la figure 2.3, on présente la fonction de covariance de Matérn avec différentes valeurs de σ, κ, ν. La figure 2.4 illustre des réalisations de processus gaussiens de moyenne 0 et qui utilisent les fonctions de covariance de Matérn de la figure 2.3, sur l'espace $D = [0,1] \times [0,1]$.

L'effet des paramètres κ et ν est très visible. En gardant κ fixe, donc en regardant les trois réalisations à gauche ou à droite, on remarque qu'en augmentant ν, le processus devient de plus en plus localisé. En gardant maintenant ν fixe, il semble qu'un plus petit κ rende le processus plus lisse. D'après l'équation 2.3, on déduit que lorsque ν est fixe

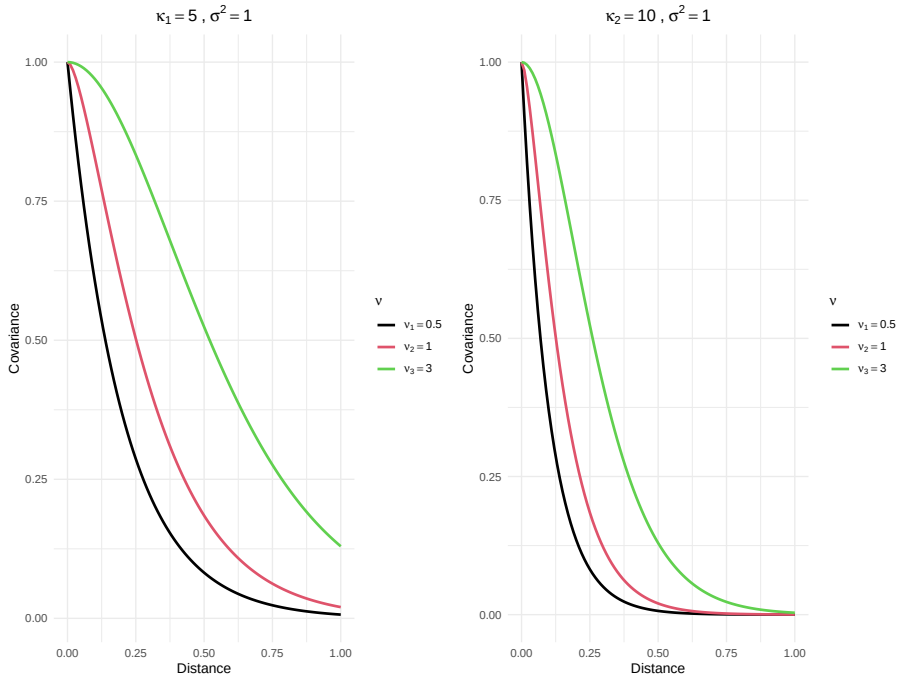

FIGURE 2.3 – Fonctions de covariance de Matérn selon différents paramètres.

et que κ augmente, alors la portée ρ diminue. On rappelle que la portée représente la distance à partir de laquelle la corrélation entre deux points est négligeable. En d'autres termes, lorsque κ augmente, la corrélation entre deux points devient négligeable plus rapidement. Finalement, on voit effectivement que déjà à $\nu = 3$, le processus engendré est très lisse. On comprend donc que lorsque $\nu \to \infty$, soit lorsque la fonction de covariance est exponentielle quadratique, le processus engendré peut souvent être considéré trop lisse.

Approximation du processus gaussien

On souhaite calculer un processus gaussien afin de capturer efficacement les dépendances spatiales. Cependant, calculer toutes les valeurs d'un processus gaussien devient rapidement coûteux en temps de calcul. L'approximation numérique devient alors essentielle pour réduire la complexité tout en préservant une représentation fidèle des phénomènes spatiaux sous-jacents. Quelques méthodes existent pour approximer numériquement un processus gaussien $Z(s)$. Une méthode, populaire il y a quelques années, consiste à partitionner la surface à l'étude en une grille relativement fine et considérer le nombre

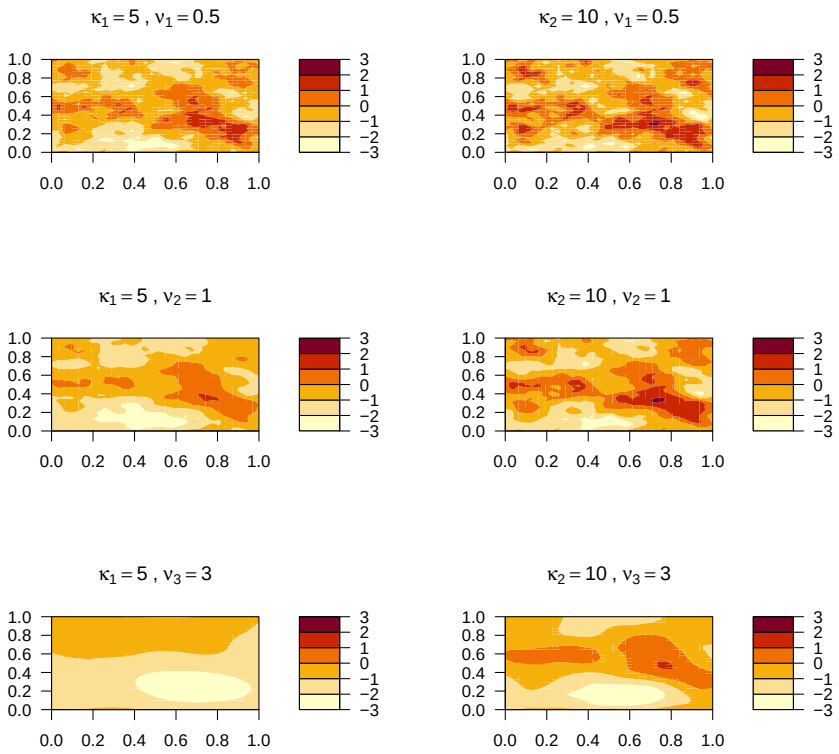

FIGURE 2.4 – Réalisations de processus gaussiens de moyennes 0 et de fonctions de covariance correspondantes aux covariance de Matérn de la figure 2.3.

de points dans chacune des cellules. Plus cette grille devient fine, plus l'approximation se rapproche de la réalité, mais cela rend la matrice de covariance du processus gaussien plus dense, et la computation plus coûteuse (MORAGA, 2023).

Une autre technique (SIMPSON et al., 2016) est beaucoup plus efficace sur le plan computationnel. De plus, cette technique considère la localisation exacte des points plutôt que simplement leur emplacement dans une grille. Cette technique utilise le résultat montré dans WHITTLE, 1963 selon lequel un processus gaussien $Z(s)$ est une solution de l'équation partielle différentielle stochastique (EPDS) suivante :

$$\tau(\kappa^2 - \Delta)^{\alpha/2} Z(s) = \varepsilon(s). \tag{2.6}$$

On retrouve le paramètre d'échelle κ défini à l'équation 2.2. Le paramètre de lissage ν fait indirectement partie de l'équation puisque $\alpha = \nu + d/2$, où d est la dimension, donc

deux dans notre cas. On fixe souvent ν à 1 puisqu'il est particulièrement difficile à estimer (WIKLE, ZAMMIT-MANGION et CRESSIE, 2019). Par conséquent, α est souvent égal à 2. Rappelons d'ailleurs que la portée $\rho = \sqrt{8\nu}/\kappa$ représente la distance entre deux points à laquelle la corrélation est négligeable. Le paramètre τ contrôle la variance σ^2 de la fonction de covariance de Matérn (2.5). Cette variance elle-même est définie par :

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}. \tag{2.7}$$

Le paramètre Δ de l'équation 2.6 représente l'opérateur laplacien $\delta^2 Z(s)/\delta x^2 + \delta^2 Z(s)/\delta y^2$, où x et y sont les coordonnées du point s en deux dimensions . Finalement, $\varepsilon(s)$ est un processus spatial à bruit blanc, composé de variables aléatoires gaussiennes non-corrélées à variance égale.

Puisqu'un processus gaussien $Z(s)$ à covariance de Matérn est une solution au EPDS décrit à l'équation(2.6) (WIKLE, ZAMMIT-MANGION et CRESSIE, 2019), il est possible d'approximer cette solution avec la méthode des éléments finis. Pour ce faire, on recouvre la surface d'une triangulation de Delaunay et on approxime le processus gaussien à covariance de Matérn, $Z(s)$, par :

$$Z(s) = \sum_{i=1}^{n} z_i \phi_i(s), \tag{2.8}$$

où n est le nombre de nœuds de la triangulation, $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ est un vecteur aléatoire gaussien multivarié, $\{\phi_i(s)\}_{i=1}^n$ est un ensemble de fonctions de base déterministes linéairement indépendantes et s est simplement un point dans la surface d'étude. Ci-dessous, nous décrirons plus en profondeur ces éléments.

La première étape pour estimer $Z(s)$ est de construire une triangulation de Delaunay (WIKLE, ZAMMIT-MANGION et CRESSIE, 2019). Une triangulation consiste simplement à décomposer un polygone en un ensemble de triangles. La triangulation d'un processus ponctuel est une triangulation où tous les points du processus sont des sommets de triangles. On appelle ces sommets des nœuds. Une triangulation de Delaunay est une triangulation d'un processus ponctuel où le plus petit angle de chaque triangle est maximisé. Les triangles allongés y sont donc défavorisés.

Dans l'analyse de MPS, il est préférable que les points spatiaux ne soient pas utilisés comme nœuds de la triangulation. En effet, il est plus habituel de générer une triangulation qui recouvre simplement la surface étudiée, avec des nœuds placés selon certains paramètres, comme la longueur maximale des arêtes des triangles, ou bien la distance minimale entre deux nœuds. On dit alors que la triangulation construite est une triangulation de Delaunay avec contraintes. L'emplacement de ces nœuds devra néanmoins être paramétrisé avec soin. La triangulation doit être assez dense pour bien estimer le processus spatial, mais une triangulation trop dense sera coûteuse computationnellement. De plus, la triangulation doit recouvrir toute la surface d'étude, avec une petite extension autour de celle-ci pour réduire les effets de bords (WIKLE, ZAMMIT-MANGION et CRESSIE, 2019). La figure 2.5 montre une triangulation de Delaunay avec contraintes recouvrant le motif simulé provenant d'un processus de Poisson non-homogène (PPNH) de la figure 2.9. Remarquons que les nœuds de la triangulation, donc les sommets des triangles en noir, ne correspondent pas aux points spatiaux en bleu et qu'il y a une petite extension autour de la surface d'étude où les triangles sont plus grands.

Concentrons-nous maintenant sur les facteurs z_i et $\phi_i(s)$ de l'équation (2.8). Tel que mentionné précédemment, $\boldsymbol{z}$ est un vecteur aléatoire gaussien multivarié défini aux n nœuds de la triangulation et sa distribution conjointe suit une $\mathscr{MN}(0, K(\tau, \kappa))$, où $K(\tau, \kappa)$ est une matrice de covariance de Matérn. Cette distribution est choisie de manière à ce que $Z(s)$ soit approximativement une solution du EPDS de l'équation(2.6) aux nœuds de la triangulation, ou en d'autres mots, approximativement un processus gaussien à covariance de Matérn aux nœuds de la triangulation.

Il ne reste maintenant qu'à interpoler cette solution au reste de la surface. C'est ici que les facteurs $\phi_i(s)$ entrent en jeu. On a déjà dit que les $\{\phi_i(s)\}_{i=1}^n$ étaient un ensemble de fonctions de base déterministes linéairement indépendantes. $\phi_i(\cdot)$ prend la valeur de 1 au nœud i, et décroît linéairement jusqu'à atteindre 0 aux nœuds voisins. Les fonctions de base de deux nœuds sont illustrées à la figure 2.6. Notons que les fonctions de bases de deux nœuds voisins se chevauchent.

Finalement, en reprenant l'équation (2.8), on peut approximer la valeur de n'importe

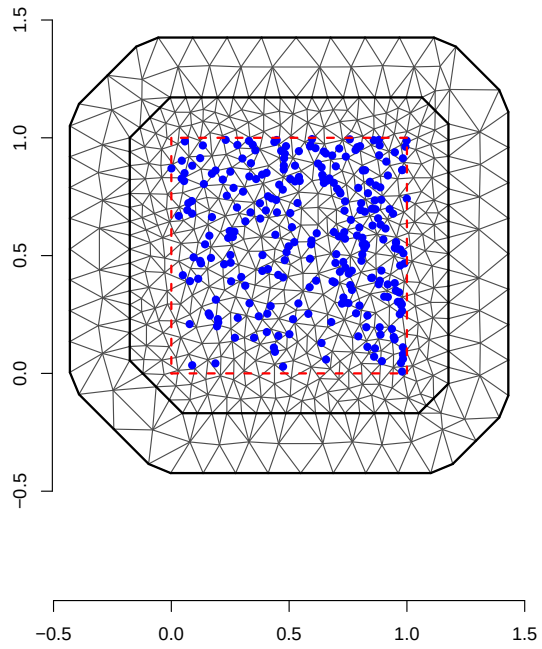

FIGURE 2.5 – Triangulation recouvrant la surface étudiée. Les points en bleu sont ceux provenant du PPNH de la figure 2.9. En pointillé rouge, ce sont les extrémités de la surface étudiée.

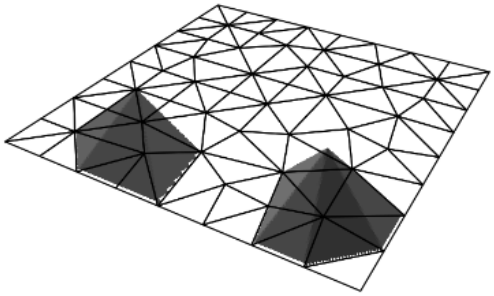

FIGURE 2.6 – Triangulation recouvrant une surface et des fonctions de bases $\phi_i(\cdot)$ de deux nœuds en foncé. Figure tirée de KRAINSKI et al., 2020

quel point s du processus gaussien $Z(s)$. D'abord, tous les $z_i, i = 1, \ldots, n$, ou n est le nombre de nœuds , sont des approximations du processus gaussien à covariance de Matérn au nœud i. Si s est directement sur le nœud i de la triangulation, la fonction de base $\phi_i(s)$ vaut alors 1, et toutes les autres fonctions de bases $\phi_{j\neq i}(s)$ valent 0. Par conséquent $Z(s)$ au nœud i vaut simplement z_i. Par contre, si s n'est pas directement sur un nœud, ce point tombe alors entre trois nœuds. Pour simplifier la notation, disons qu'il s'agit des nœuds 1, 2 et 3. Cette situation est illustrée à la figure 2.7. Dans cet exemple, les valeurs du processus gaussien aléatoire z_1, z_2, z_3 sont connues aux nœuds 1, 2 et 3. Les fonctions de bases $\phi_1(s), \phi_2(s), \phi_3(s)$ valent 1 à leur nœuds respectif et décroissent linéairement jusqu'à 0 aux nœuds voisins. Elles sont donc plus grandes que 0 (mais plus petites que 1) au point s. Toutes les autres fonctions de base (différentes de 1, 2 et 3) valent 0 en s. On peut finalement approximer $Z(s)$ de la manière suivante :

$$Z(s) \approx \frac{T_1}{T} z_1 + \frac{T_2}{T} z_2 + \frac{T_3}{T} z_3, \tag{2.9}$$

où T_i est l'aire des sous-triangles 1, 2 et 3, et T, l'aire du triangle global. Notons que les $\frac{T_i}{T}$ sont les ϕ_i.

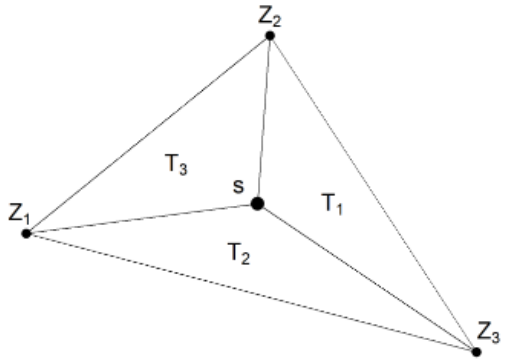

FIGURE 2.7 – Le point spatial s entre les nœuds 1, 2 et 3. Les arêtes entre les nœuds 1, 2 et 3 forment un triangle d'une triangulation. T_1, T_2 et T_3 représentent les aires des petits triangles. Figure tirée de MORAGA, 2023.

2.3 Processus spatiaux pour motifs ponctuels spatiaux (MPS)

Dans cette recherche, nous utiliserons des données ponctuelles décrivant la localisation géographique des accidents routiers sur l'île de Montréal. Ces données ponctuelles forment des MPS, et nous voudrons éventuellement établir comment la localisation des accidents a évolué depuis le début de la pandémie. Cette section introduira plusieurs concepts de l'analyse de MPS ainsi que des modèles correspondants.

Introduisons d'abord la notation. Les réalisations d'un processus ponctuel spatial (PPS) prennent la forme de points spatiaux se situant dans une région plane $A \subset \mathbb{R}^2$. On dénote $N(A)$ la variable aléatoire représentant le nombre de points spatiaux dans la région plane A. Une réalisation d'un processus ponctuel spatial est constituée d'un ensemble dénombrable de points $\{s_1, s_2, \ldots, s_{N(A)}\}$, où le nombre de points $N(A)$ peut être différent d'une réalisation à l'autre.

Attardons-nous maintenant à certains concepts. On décrit habituellement un MPS par sa densité spatiale, qui possède les mêmes propriétés que la fonction de densité d'une variable aléatoire (BIVAND, E. J. PEBESMA et GOMEZ-RUBIO, 2008a). La fonction de densité spatiale, notée $f(s)$, décrit la probabilité d'observer un évènement au point s. Elle a pour domaine la région d'étude A.

L'intensité d'un processus spatial, notée $\lambda(s)$, est une autre fonction qui caractérise la distribution spatiale. Cette fonction fournit le nombre d'évènements attendus par unité de surface au point s. Notons que l'intensité est proportionnelle à la densité spatiale. En d'autres mots, deux MPS peuvent avoir la même densité spatiale, mais celui avec le plus de points aura une plus grande intensité. Ces concepts seront approfondis dans cette section.

2.3.1 Processus de Poisson

Le processus de Poisson est un processus ponctuel spatial. Il est utilisé pour modéliser des MPS. Un processus de Poisson avec une intensité $\lambda(\cdot)$ a les propriétés suivantes (MORAGA, 2023) :

1. $N(A)$, le nombre d'évènements sur une surface A suit une distribution de Poisson avec moyenne $\mu(A) = \int_A \lambda(x)dx$. En termes mathématiques, on a :

$$P(N(A) = n) = \frac{e^{-\mu(A)} \cdot \mu(A)^n}{n!} \tag{2.10}$$

2. Sachant que $N(A) = n$, les localisations des évènements sur A sont indépendantes et identiquement distribuées (iid) avec une fonction de densité proportionnelle à leur intensité $\lambda(\cdot)$.

Un processus de Poisson peut être homogène ou non-homogène. C'est-à-dire que l'intensité $\lambda(\cdot)$ peut être soit constante ou varier dans l'espace. Dans les deux cas, les évènements sont indépendants les uns des autres et distribués selon l'intensité. Avant de développer sur les processus de Poisson homogène et non-homogène, introduisons l'intensité plus formellement.

2.3.2 Intensité

L'intensité d'un processus ponctuel spatial est le nombre de points (évènements) attendus par unité de surface. C'est en quelque sorte l'équivalent de la moyenne lorsqu'on travaille avec des données numériques (BADDELEY, RUBAK et TURNER, 2015). L'intensité peut être constante, ou varier spatialement.

Posons dx, une petite région contenant le point x, et $|dx|$, l'aire de la région dx, la fonction d'intensité de premier ordre est définie comme suit (MORAGA, 2023) :

$$\lambda(x) = \lim_{|dx| \to 0} \frac{E[N(dx)]}{|dx|}. \tag{2.11}$$

Posons maintenant dy, une petite région contenant le point y, et $|dy|$, l'aire de la région dy, alors la fonction d'intensité de second ordre est

$$\lambda_2(x,y) = \lim_{|dx|\to 0, |dy|\to 0} \frac{E[N(dx)N(dy)]}{|dx||dy|}. \tag{2.12}$$

La fonction d'intensité de second ordre est en quelque sorte l'équivalent en statistique spatiale de la covariance en statistique classique (BADDELEY, RUBAK et TURNER, 2015). En d'autres mots, elle permet de mesurer l'interaction entre deux points d'un processus ponctuel. La fonction K, dont il sera question plus loin, est une manière alternative de mesurer les liens de second ordre.

2.3.3 Processus de Poisson homogène (PPH)

Le processus de Poisson homogène est souvent utilisé comme modèle de référence pour les motifs spatiaux aléatoires (BADDELEY, RUBAK et TURNER, 2015). En effet, il est caractérisé par des évènements indépendants et distribués uniformément. Un PPH pourrait être utilisé pour modéliser des nids de poules dans une rue, ou bien les raisins dans une miche de pain. En analyse spatiale, on cherche généralement à expliquer des regroupements de points, ou de trouver une tendance qui explique le motif dans les points. Par conséquent, on cherche généralement à s'éloigner du modèle de PPH.

Comme un PPH a une intensité homogène, $\lambda(x) = \lambda$ pour tout $x \in A$, on peut réduire les équations 2.11 et 2.12 et obtenir les fonctions d'intensité du premier ordre et du second ordre suivantes :

$$\begin{aligned} \lambda(x) = \lambda &= \frac{E[N(A)]}{|A|}, \\ \lambda_2(x,y) &= \lambda_2(||x-y||) = \lambda_2(h), \end{aligned} \tag{2.13}$$

où $h = ||x-y||$ est la distance entre les points x et y. On remarque que l'intensité ne dépend que de la distance h entre les deux points. On peut donc en déduire que le PPH est stationnaire.

On qualifie parfois le PPH de processus spatial complètement aléatoire (CSR, pour *Complete Spatial Randomness*). Le processus CSR est un processus spatial où les évè-

nements ont une probabilité égale de se produire n'importe où sur la surface et où ils sont indépendants les uns des autres (DIGGLE, 2013). En fait, le PPH est une définition formelle d'un processus CSR.

Notons qu'un PPH ne génère pas de motif régulier, soit un motif où les distances entre les points sont à peu près régulières et ordonnées. En effet, une distribution uniforme des points implique un certain aspect aléatoire dans le motif. La figure 2.8 montre un motif régulier et un motif issu d'un PPH. Remarquons que dans le motif issu d'un PPH, on peut voir que le nombre de points générés est un peu différent du nombre de points attendus.

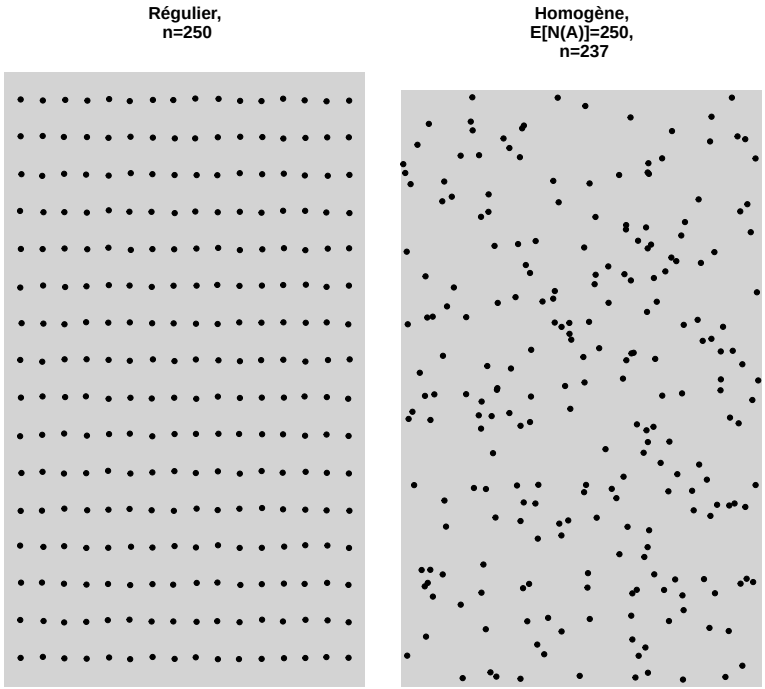

FIGURE 2.8 – À gauche : Motif spatial régulier. À droite : Motif spatial généré d'un PPH.

2.3.4 Processus de Poisson non-homogène (PPNH)

Il s'agit du modèle le plus intéressant pour de nombreuses applications. En effet, dans la plupart des cas, il n'est pas réaliste de supposer qu'un processus ponctuel est homogène. Le PPNH est une généralisation du PPH où l'on permet à l'intensité de varier dans l'espace. Un PPNH est un processus non-stationnaire. La figure 2.9 montre d'ailleurs deux

motifs, l'un provenant d'un PPH et l'autre d'un PPNH. Encore une fois que le nombre de points observés est différent du nombre de points attendus.

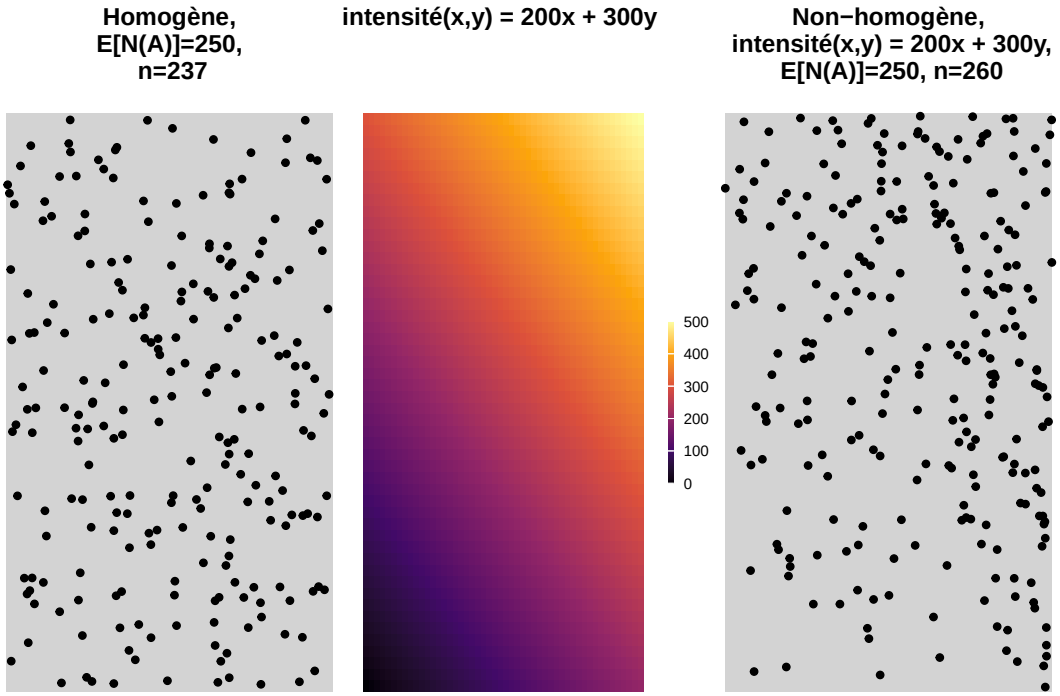

FIGURE 2.9 – Gauche : Motif spatial ponctuel généré par un PPH, où E[N(A)] = 250, et 237 points sont générés (le même qu'à la figure 2.8). Milieu : Intensité qui génère le motif ponctuel spatial de droite. Droite : Motif ponctuel spatial généré par un processus de Poisson non-homogène d'intensité illustré au milieu, où E[N(A)] = 250 ici aussi, et où 260 points sont générés.

Un PPNH peut être utilisé pour modéliser des évènements indépendants dont la localisation n'est pas uniforme. Le nombre de crimes dans une ville pourrait être modélisé par un PPNH, puisque ce nombre varie selon le quartier et que les crimes sont généralement indépendants les uns des autres.

Estimation d'une intensité non-homogène

L'estimation de l'intensité d'un motif provenant d'un processus non-homogène ou non-stationnaire est plus complexe que celle d'un motif provenant d'un processus homogène. Une méthode courante est d'estimer la densité à l'aide d'une fonction noyau

(GONZALEZ et MORAGA, 2023). On estime la fonction de densité $f(s)$ définissant la vraisemblance d'observer un évènement à une localisation s. Cette fonction de densité s'intègre nécessairement à 1 sur la région d'étude. La fonction d'intensité $\lambda(s)$ fournit le nombre attendu d'évènements par unité à la localisation s et est donc proportionnelle à la fonction de densité :

$$\lambda(s) = f(s) \int_A \lambda(u)du = f(s) \times n,$$

où $\int_A \lambda(u)du = n$ est la constante proportionnelle représentant le nombre de points sur la région d'étude. On peut estimer la fonction de densité au point s de la manière suivante :

$$\hat{f}(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^2} K\left(\frac{s_i - s}{h}\right),$$

où s_i représente les observations, n est le nombre de points, h est un paramètre de lissage appelé bande passante et la fonction de noyau $K(\cdot)$ (*kernel* en anglais) est une fonction symétrique telle que $K(s) \geq 0$ pour tout s et $\int_A K(u)du = 1$. Notons que la fonction $K(\cdot)$ n'a aucun lien avec la fonction K de Ripley. Des exemples de fonctions de noyau sont la fonction Epanechnikov, gaussienne, disque ou quartique, qui sont illustrées à la figure 2.10. On remarque que la bande passante a un bien plus grand effet sur l'estimation de l'intensité que la fonction de noyau. En fin de compte, étant donné la relation étroite entre la fonction de densité et d'intensité (facteur de n), on peut illustrer l'une ou l'autre. Nous avons choisi d'illustrer l'intensité à la figure 2.10. Le choix de la bande passante est assez subjectif et habituellement fait pour donner une intensité lisse (MORAGA, 2023).

À proximité des bordures de la région d'étude, l'estimation par noyau tend à être faussée puisque pour une même bande passante (le diamètre autour de s où on cherche les voisins s_i), on a nécessairement moins de voisins. Il faut alors ajuster l'estimation en divisant par la surface qui fait réellement partie de la région d'étude. On divise alors par le terme suivant :

$$\int_A h^{-2} K\left(\frac{\mu - s}{h}\right) d\mu. \tag{2.14}$$

Dans la figure 2.10, cet ajustement est déjà apporté.

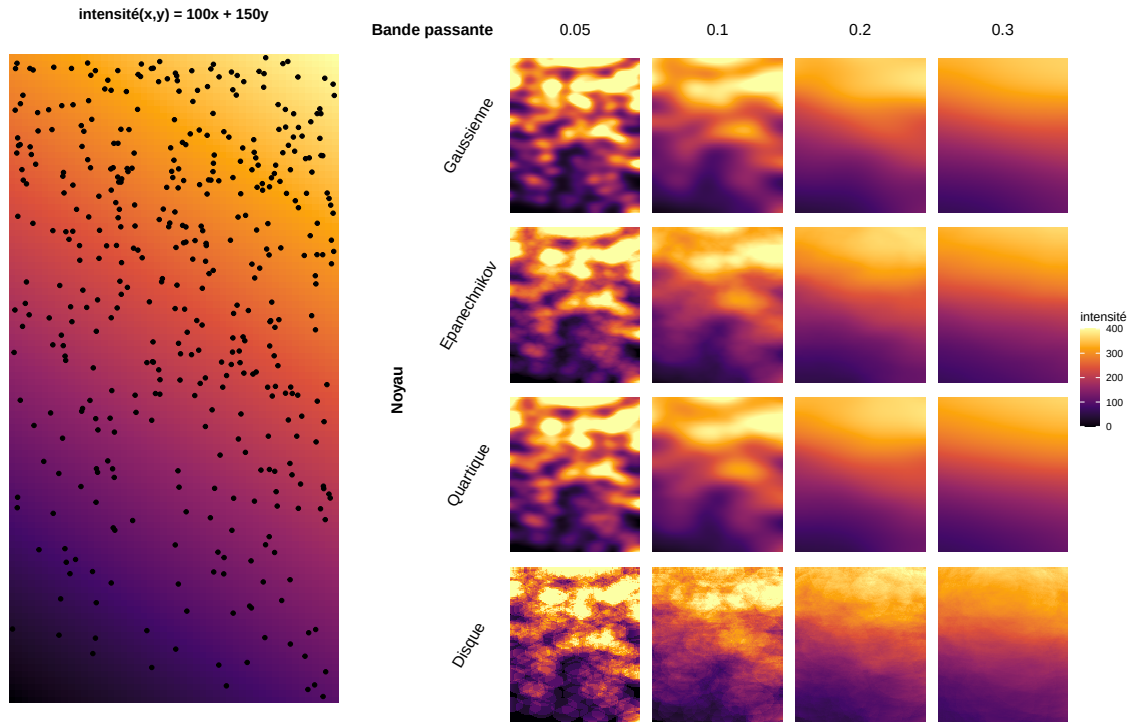

FIGURE 2.10 – Intensité selon différents noyaux. À gauche : représentation de l'intensité non-homogène à partir de laquelle est simulée le MPS en noir. À droite : l'intensité estimée avec différentes bandes passantes et fonctions de noyau. L'échelle de couleur est la même pour tous les graphiques.

2.3.5 Test d'homogénéité

Tel que mentionné précédemment, l'un des objectifs de la modélisation d'un MPS est de se distancer du PPH (ou processus CSR pour *Complete Spatial Randomness*). Deux tests statistiques permettant de vérifier si tel est le cas seront présentés dans cette section. Ces deux tests statistiques sont la méthode des quadrants et de la fonction K. On cherche à y tester l'hypothèse nulle H_0 selon laquelle le MPS est issu d'un processus CSR. L'hypothèse alternative H_1 est simplement que le motif n'est pas issu d'un processus CSR.

Méthode des quadrants

La méthode des quadrants est une méthode simple où l'on partitionne la surface étudiée en m quadrants d'aire égale. Le nombre d'évènements se déroulant dans les quadrants suit alors des distributions de Poisson iid. Il est donc naturel d'utiliser le test χ^2 de Pear-

son, avec la statistique de test suivante :

$$X^2 = \sum_{j=1}^{m} \frac{(n_j - \frac{n}{m})^2}{\frac{n}{m}} \tag{2.15}$$

où n est le nombre de points sur la surface complète, m, le nombre de quadrants et n_j, le nombre de points dans le quadrants j, où $j = 1, \ldots, m$. Sous l'hypothèse nulle H_0, X^2 suit une loi χ^2_{m-1}. La figure 2.11 montre une partition des motifs ponctuels simulés aux figures 2.8 et 2.9.

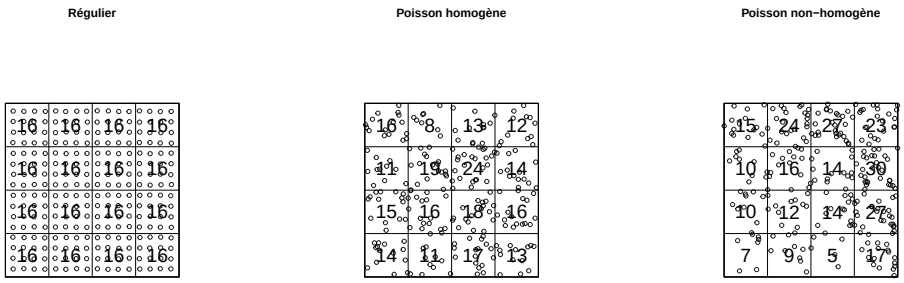

FIGURE 2.11 – MPS partitionnés en 16 quadrants. On retrouve le nombre de points dans chaque quadrant.

On obtient des statistiques de test de 7.88×10^{-30}, 24.29, 55.32 pour le motif régulier et ceux provenant d'un PPH et d'un PPNH. Les valeurs-p sont donc plus petites que 10^{-5} pour le motif régulier et PPNH et de 0.89 pour le motif provenant du PPH. On rejette donc l'hypothèse nulle pour le motif provenant du PPNH et le motif régulier. Ceci était attendu, puisque ces deux motifs ne provenaient effectivement pas d'un PPH (ou processus CSR).

Par contre, la méthode des quadrants a quelques défauts. D'abord, la puissance de ce test dépend de la dimension des quadrants (BADDELEY, RUBAK et TURNER, 2015). La puissance d'un test utilisant des quadrants trop petits ou trop grands tombe à zéro. De plus, lorsque la région étudiée n'est pas régulière, il peut être difficile de la partitionner en m quadrants de surface égale.

Méthode de la fonction K de Ripley

Tel que mentionné plus tôt, la fonction K de Ripley est une manière alternative de mesurer les propriétés du second ordre de l'intensité. Elle peut aussi être utilisée pour tester l'homogénéité d'un processus. La fonction K mesure la dépendance entre des localisations se trouvant à une distance r (BADDELEY, RUBAK et TURNER, 2015) :

$$K(r) = \lambda^{-1} E[N_0(r)],$$

où λ est la fonction d'intensité du processus spatial ponctuel et N_0 est le nombre de points voisins dans un rayon r autour d'un point arbitraire. Notons que λ est l'intensité globale de la région d'étude A. On peut facilement l'estimer avec $\hat{\lambda} = n/|A|$, où n est le nombre de points dans A, et $|A|$ est l'aire de A. Puisque l'intensité d'un PPH est constante et que l'intensité représente le nombre de points par unité de surface, le nombre de points dans rayon r autour d'un point arbitraire est $\lambda \pi r^2$. La fonction K d'un PPH est :

$$\begin{aligned} K(r) &= \lambda^{-1} \times \lambda \pi r^2 \\ &= \pi r^2, \end{aligned} \tag{2.16}$$

et ne dépend donc pas de l'intensité.

Intuitivement, un motif où les points forment des *clusters* aura un $N_0(r)$ plus grand que celui d'un motif homogène puisque la probabilité qu'un point soit proche d'autres points est plus grande. Par conséquent, il aura une plus grande valeur de fonction K. Avec un raisonnement semblable, on conclut qu'un motif régulier aura une plus petite valeur de fonction K que celle du motif issu d'un PPH.

Il ne reste maintenant qu'à estimer la fonction de K pour des processus spatiaux qui ne sont pas des PPH. DIXON, 2002 nous indique qu'on peut estimer $K(r)$ par

$$\begin{aligned} \hat{K}(r) &= \hat{\lambda} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq r) \\ &= \frac{|A|}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq r), \end{aligned} \tag{2.17}$$

où n est le nombre de points total dans la région étudiée A, $|A|$ est l'aire de A, I, une fonction indicatrice, d_{ij}, la distance entre les points i et j, et w_{ij}, un ajustement pour les effets de bords. La fonction indicatrice ne sert qu'à indiquer si un point est à une distance r ou moins d'un autre. L'ajustement pour effet de bord w_{ij} peut être estimé de plusieurs manières, mais la plus commune est celle de Ripley, où w_{ij} représente la proportion du disque de rayon d_{ij} se trouvant dans la région à l'étude A.

En reprenant les motifs ponctuels simulés aux figures 2.8 et 2.9, on obtient les fonctions K provenant d'un PPH (appelée K_{pois} dans la figure) et provenant du MPS à l'étude (appelée K_{iso} dans la figure) présentées à la figure 2.12. Notons que « iso » signifie que l'ajustement pour effet de bord utilisé est celui de Ripley. On remarque que la fonction K estimée du motif issu d'un PPH (au milieu à la figure 2.12) se confond presque avec la fonction K théorique, qui est justement censée provenir d'un PPH. On voit aussi clairement que les fonctions K des deux autres motifs sont différentes de la fonction K théorique. Il serait plus intéressant de tester cela statistiquement, donc d'avoir un intervalle de confiance autour de la courbe du K théorique. Un test de Monte-Carlo peut être utilisé pour obtenir ces intervalles de confiance, dont les étapes sont (MORAGA, 2023) :

1. Générer un grand nombre M de motifs provenant d'un PPH contenant le même nombre de points que le motif observé. M pourrait être égal à 100 ou à 1000 par exemple.
2. Pour chacun des M motifs générés, estimer la fonction $\hat{K}_1(r), \ldots, \hat{K}_M(r)$.
3. À toutes les distances r, calculer l'intervalle de confiance à 95%.

On rejette donc l'hypothèse nulle H_0, selon laquelle le motif provient d'un PPH, si la fonction K estimée (provenant du MPS à l'étude) se trouve en dehors de l'intervalle de confiance à 95%, même si ce n'est qu'à un point.

La fonction *envelope* de la bibliothèque *spatstat* de *R* réalise ces étapes. Notons que cette fonction offre la possibilité d'appliquer une transformation à la fonction K, pour permettre une interprétation visuelle plus facile. Cette transformation permet d'obtenir la

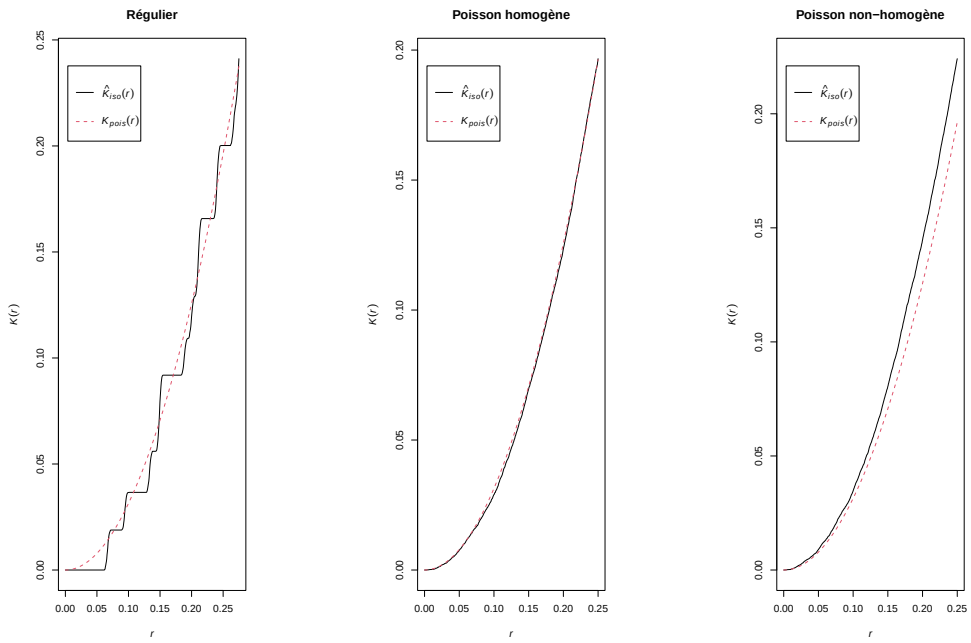

FIGURE 2.12 – Comparaison de la fonction K théorique (provenant d'un PPH) en pointillé rouge et la fonction K estimée (provenant du MPS à l'étude) en noir. Les motifs sont, de gauche à droite, régulier, issu d'un PPH et issu d'un PPNH.

fonction L :

$$L(r) = \sqrt{\frac{K(r)}{\pi}}.$$

La transformation L transforme la courbe de la fonction K en une droite. La figure 2.13 reprend les fonctions K présentées à la figure 2.12, auxquelles on a appliqué la transformation L et généré des intervalles de confiance. La fonction L du motif issu d'un PPH est effectivement la seule à rester dans l'intervalle de confiance pour toutes les distances r.

2.3.6 Processus de Cox

Les processus de Poisson sont souvent trop simples, rendant difficile la modélisation des agrégations de points. Un processus de Cox est une extension naturelle des processus de Poisson qui permet de modéliser des regroupements (*clusters*) (SCHABENBERGER et GOTWAY, 2005). Ce processus est une généralisation du PPNH, dans laquelle l'intensité est considérée comme une variable aléatoire, ce qui permet une variance de processus plus élevée que la moyenne. De plus, on rappelle que le PPNH suppose que le nombre

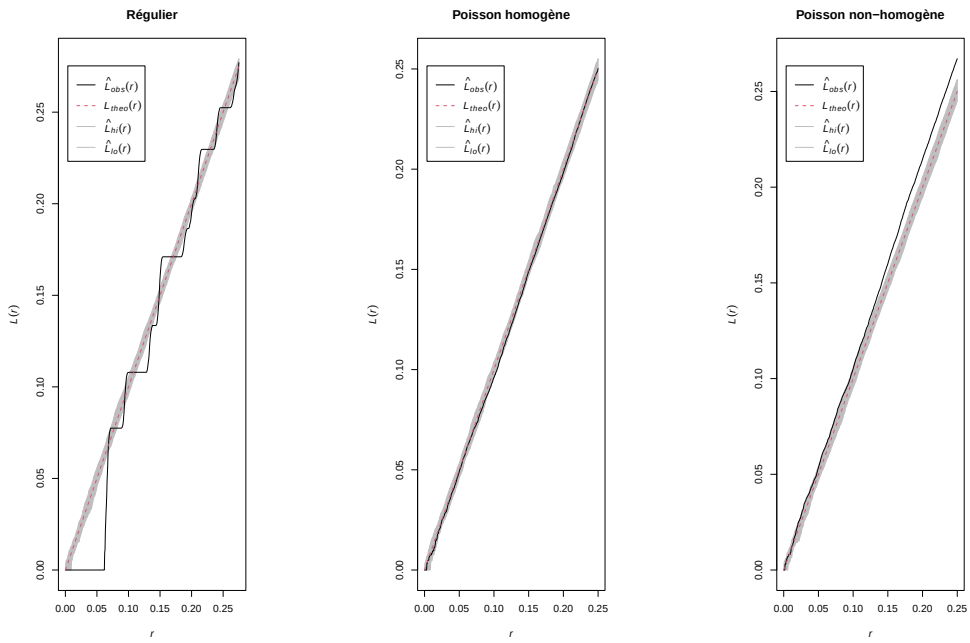

FIGURE 2.13 – Fonctions L provenant des fonctions K de la figure 2.12, avec un intervalle de confiance.

d'événements dans deux sous-espaces disjoints est indépendant. Cette hypothèse d'indépendance ne s'applique cependant pas au modèle de Cox, qui permet ainsi de modéliser des situations dans lesquelles cette hypothèse n'est pas réaliste.

Dans un processus de Cox, l'intensité est aléatoire et définie pour chaque point s par $\Lambda(s)$, dont $\lambda(s)$ est une réalisation. Si on conditionne le processus de Cox sur une réalisation de $\Lambda(s)$, le processus résultant est un processus de Poisson. Il est donc possible de modéliser des processus ponctuels stationnaires et non-stationnaires. Cependant, $\Lambda(s)$ n'est généralement pas observable et il est difficile de différencier un processus de Cox du processus de Poisson équivalent.

Un grand avantage des processus de Cox est que, même si on modélise une intensité qui varie spatialement, le processus de Cox peut rester stationnaire tout en étant influencé par des covariables. En effet, supposons une intensité $\lambda(s)$ non-homogène, on peut définir l'intensité aléatoire $\Lambda(s) = \beta X(s) + Z(s)$, où $X(s)$ est un vecteur de variables explicatives au point spatial s et β, les coefficients associés. Si la combinaison linéaire $\beta X(s)$ explique la variation de l'intensité, on dit que le processus de Cox $\Lambda(s)$ est un processus

stationnaire pondéré par l'intensité (BADDELEY, MØLLER et WAAGEPETERSEN, 2001). On peut utiliser des fonctions de covariances stationnaires comme celle de Matérn pour définir la covariance spatiale du processus gaussien $Z(s)$.

Processus log-gaussien de Cox

Le processus de Cox le plus utilisé est le processus log-gaussien de Cox (LGCP pour *Log-Gaussian Cox Process*). Il s'agit aussi d'une généralisation d'un PPNH. C'est un processus de Cox où l'intensité est modélisée par :

$$\Lambda(s) = \exp(Z(s)), \tag{2.18}$$

où $\Lambda(s)$,est une variable aléatoire représentant l'intensité, et $Z = \{Z(s) : s \in \mathbb{R}^2\}$ est un processus gaussien. Les LGCP sont généralement utilisés pour modéliser des phénomènes induits par l'environnement, comme la distribution spatiale de malades non-contagieux (DIGGLE et al., 2013). Le motif (*pattern*) d'individus malades observé peut alors être expliqué par une variation spatiale dans l'exposition à la maladie, ainsi qu'à d'autres facteurs de risque. La distribution spatiale de malades contagieux est, entre autre, le résultat de la transmission d'une personne à l'autre. Il est difficile de distinguer empiriquement si un point résulte d'une interaction avec un autre point (par contagion) ou s'il est le résultat de son environnement. Il faut s'assurer que la localisation des points n'est pas directement induite par l'interaction entre ceux-ci. Les accidents ne sont généralement pas induits par d'autres accidents.

Dans un LGCP, conditionnellement à la réalisation de $\Lambda(s)$, notée $\lambda(s)$, on obtient un PPNH d'intensité $\lambda(s)$. On retrouve donc les propriétés d'un PPNH, soit que le nombre d'évènements suit une distribution Poisson, $\mu(A) = \int_A \Lambda(s)ds$, et que la localisation des évènements sur A est indépendante et identiquement distribuée.

On est habituellement intéressé par l'effet des covariables spatiales sur l'intensité. Une manière d'ajouter ces effets est de reparamétriser le processus de la manière suivante :

$$\Lambda(s) = \exp(\beta X(s) + Z(s)), \tag{2.19}$$

où $X(s)$ est un vecteur de covariables, β, le vecteur des coefficients associés, et $Z(s)$, le processus gaussien représentant la composante spatiale. $Z(s)$ est généralement modélisé avec une moyenne de zéro et une fonction de covariance de Matérn. On rappelle que la fonction de covariance de Matérn est utilisée dans l'analyse spatiale et spatio-temporelle pour sa grande flexibilité. De plus, puisqu'elle ne dépend que de la distance entre deux points et non de leur localisation exacte, elle n'est pas trop coûteuse computationnellement.

2.4 Implémentation en R

Heureusement, l'implémentation d'un modèle LGCP se fait relativement bien sur *R*. La librairie *INLA* permet d'ajuster une approximation des modèles LGCP dans un contexte bayésien. De la construction de la triangulation, à l'utilisation d'un modèle EPDS pour approximer le processus gaussien, jusqu'à l'ajout des covariables, tout peut y être fait. Cependant, comme *INLA* offre énormément de flexibilité et d'options, une bonne connaissance de la bibliothèque et une bonne compréhension théorique des modèles sont nécessaires. Des exemples de modélisation de LGCP à l'aide d'*INLA* peuvent être trouvés dans MORAGA, 2023 et GÓMEZ-RUBIO, 2020.

La bibliothèque *inlabru* quant à elle facilite l'utilisation d'*INLA* pour la modélisation spatiale. Cette bibliothèque dispose d'une fonction `lgcp`$(\cdot)$ permettant de plus facilement implémenter ce modèle. Bien entendu, on obtient les mêmes résultats en utilisant *INLA* et *inlabru*. Notons qu'*inlabru* est développé par des auteurs qui apparaissent tout au long de cette section tels que Finn Lindgren. Des exemples de modélisation à l'aide d'*inlabru* peuvent être trouvés dans LINDGRENN et BORCHERS, 2024.

Il faut tout de même s'occuper de quelques hyperparamètres. En particulier, ceux permettant d'ajuster le modèle EPDS. Dans l'équation (2.6), les trois paramètres α, κ et τ ont un rôle à jouer. On rappelle qu'α est généralement fixé à 2 puisque les paramètres sont difficilement identifiables, que $\kappa = \sqrt{8\nu}/\rho$ est lié à la portée ρ, et finalement que τ est proportionnel à σ^{-1} (voir équation 2.7).FUGLSTAD et al., 2019 propose des *a priori*s

de pénalisation de la complexité pour les paramètres σ et ρ. Ces deux paramètres sont interprétables et ont des propriétés théoriques intéressantes (SIMPSON et al., 2016). Les distributions *a priori* sont spécifiées de manière à définir indirectement les hyperparamètres à travers les relations suivantes :

$$P(\rho < \rho_0) = p_\rho, P(\sigma > \sigma_0) = p_\sigma, \tag{2.20}$$

où ρ_0 et p_ρ sont le quantile de la queue inférieure et la probabilité de la portée, alors que σ_0 et p_σ sont le quantile de la queue supérieure et la probabilité de l'écart-type du processus gaussien de covariance de Matérn.

Un autre grand avantage de l'utilisation de la modélisation bayésienne est qu'on peut comparer des modèles entre eux. Le DIC est un critère populaire dans le choix de modèles. Le DIC est similaire à l'AIC (GÓMEZ-RUBIO, 2020). Plus le DIC est petit, mieux le modèle s'ajuste aux données.

Notons que le calcul du DIC d'*INLA* et d'*inlabru* ne concordent pas toujours. Finn Lindgren, développeur principal d'*inlabru* et chercheur important dans la modélisation de MPS, mentionne que cette différence provient des multiples manières de construire la vraisemblance d'un MPS. Il continue en disant que les exemples qui utilisent la technique décrite dans MORAGA, 2023 et GÓMEZ-RUBIO, 2020, donc qui construisent le modèle LGCP dans *INLA*, n'utilisent pas la même approximation de vraisemblance que celle décrite dans SIMPSON et al., 2016. Dans le reste de ce mémoire, lorsqu'il sera question de DIC de MPS, ce sera celui provenant d'*inlabru*.

Finalement, la figure 2.14 montre la moyenne *a posteriori* de l'intensité calculé à l'aide d'un LGCP.

2.4.1 Modèle spatio-temporel

Étendons maintenant le modèle LGCP pour y ajouter un effet temporel. Le LGCP spatio-temporel est défini par :

$$\Lambda(s,t) = \exp(\beta X(s,t) + Z(s,t)), \tag{2.21}$$

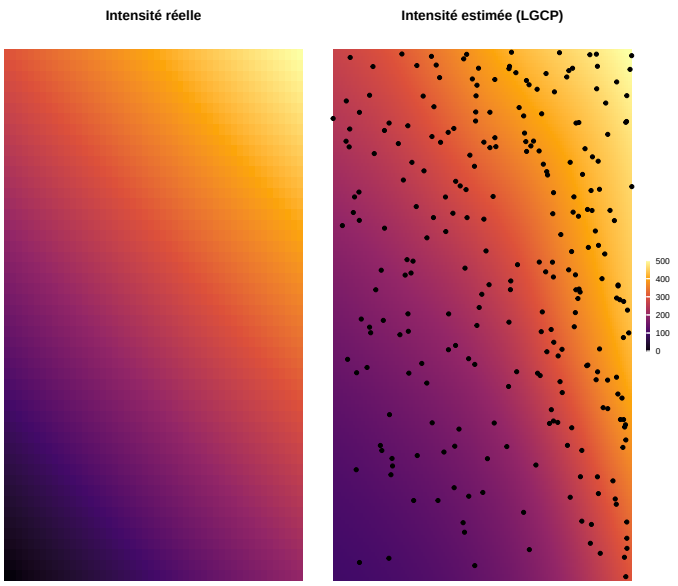

FIGURE 2.14 – Comparaison entre l'intensité réelle à gauche et l'intensité approximée par le LGCP à droite. Les points en noir sont ceux du MPS de la figure 2.9.

soit le même modèle qu'on avait, auquel on ajoute des indices temporels t. On choisit alors la covariance pour qu'elle soit séparable, c'est-à-dire que $\text{Cov}(s_1, s_2, t_1, t_2) = \text{Cov}(s_1, s_2) \times \text{Cov}(t_1, t_2)$. Cela rend la computation moins coûteuse, mais implique également qu'il n'y a aucune interaction entre le temps et l'espace, ce qui n'est pas idéal dans le cas des accidents de la route, où des effets spatio-temporels complexes peuvent exister. En cas de nécessité, des modèles plus complexes permettant de prendre en compte ces interactions pourraient être envisagés.

La corrélation spatiale reste un processus gaussien de covariance Matérn, qu'on approxime avec un EPDS. La corrélation temporelle est indépendante de la corrélation spatiale. On modélise la covariance entre deux observations dans le temps, à un point spatial s en particulier. On modélise souvent la matrice de covariance temporelle à l'aide d'une structure autorégressive de type 1 (AR1) ou d'une structure d'équicorrélation. Dans une structure AR1, la corrélation entre deux observations décroît géométriquement avec le temps écoulé. On peut l'exprimer comme

$$Z(t) = \phi Z(t-1) + \varepsilon_t, \tag{2.22}$$

où ε_t est une erreur gaussienne indépendante, et ϕ, un paramètre qui contrôle la corrélation entre les observations consécutives dans le temps. Notons que la notation habituelle de ϕ est ρ, mais qu'on ajuste pour éviter toute confusion avec la portée. Plus ϕ est proche de 1, plus la corrélation temporelle est forte. De plus, la corrélation temporelle n'est paramétrée que par ϕ. La structure d'équicorrélation suppose quant à elle que toutes les paires d'observations ont la même corrélation, peu importe le temps s'étant écoulé entre les deux. Il est encore une fois possible d'implémenter ces modèles sur *INLA* et *inlabru*. Le modèle spatio-temporel par défaut sur *inlabru* utilise une structure d'équicorrélation pour la corrélation temporelle. Les *a priori* utilisés sont ceux par défaut sur *INLA*, décrits ici INLA, s. d.

2.4.2 Comparaison des intensités

Il est également intéressant de modéliser d'autres marques des accidents que le temps, comme par exemple, la gravité de l'accident. Dans cette dernière section sur les MPS, nous présenterons une technique naïve pour comparer deux intensités, ainsi qu'un modèle permettant de modéliser deux MPS simultanément, comme un motif pour les accidents graves, et un autre pour les accidents mortels.

Ratio d'intensité

Cette technique fait simplement le ratio des deux intensités provenant de MPS dans une même région. Le ratio d'intensité ι de deux intensités notées λ_0 et λ_1 prend la forme suivante :

$$\iota(s) = \alpha \frac{\lambda_0(s)}{\lambda_1(s)}. \tag{2.23}$$

Le facteur constant d'ajustement α peut simplement être estimé en faisant un ratio du nombre de points observés dans un MPS et dans l'autre de cette manière

$$\alpha = \frac{n_1}{n_0}.$$

n_i est le nombre d'évènements provenant du ie motif spatial. Bref, le ratio d'intensité permet de comparer des intensités, qu'elles soient estimées ou non, mais il s'agit d'un outil visuel et non d'un test statistique.

Modèle LGCP multivarié

Un modèle LGCP multivarié modélise simultanément plusieurs MPS. Voici un modèle proposé dans DIGGLE et al., 2013 :

$$\Lambda_k(s) = exp\left(\beta_k + S_0(s) + S_k(s)\right), \tag{2.24}$$

où β_k est une ordonnée à l'origine spécifique au groupe k, S_0 est un processus gaussien commun à tous les points et S_k est un processus gaussien au points de type k. Ce genre de modèle permet de modéliser des marques. On les implémente à l'aide de *INLA* ou *inlabru* en définissant des objets de vraisemblance distincts pour chaque motif ponctuel spatial (MPS) que l'on souhaite modéliser. Un exemple de ce type de modèle peut être trouvé à BACHL, 2024.

Dans ce mémoire, nous utiliserons des modèles LGCP multivarié avec des covariables mais une seule ordonnée à l'origine, et deux processus spatiaux par MPS. Par exemple, nous voudrons modéliser la différence entre les accidents pré et pendant la COVID-19. Pour ce faire nous utiliserons deux vraisemblances, une pour chaque processus S :

$$\begin{aligned} \Lambda_{\text{Pré}}(s) &= \exp\left(\beta_0 + \beta X + S_{\text{Commun}}(s) - 0.5 * S_{\text{Différent}}(s)\right), \\ \Lambda_{\text{COVID}}(s) &= \exp\left(\beta_0 + \beta X + S_{\text{Commun}}(s) + 0.5 * S_{\text{Différent}}(s)\right). \end{aligned} \tag{2.25}$$

Cette paramétrisation des processus gaussiens S provient directement de BACHL, 2024. Ces deux vraisemblances sont ensuite modélisées simultanément, ce qui permet de véritablement observer un processus spatial de la différence entre les deux, sans favoriser l'un des deux cas. Il est possible d'ajouter des effets temporels, mais ces modèles deviennent alors computationnellement très coûteux.

Pour conclure, l'analyse des MPS évolue énormément depuis quelques années et les outils deviennent de plus en plus accessibles. Dans cette section, nous avons présenté les

bases de l'analyse des MPS et nous avons continué vers des modèles complexes permettant d'inclure des marques et un effet temporel.

2.5 Analyse de données surfaciques (zonales)

Les données surfaciques sont souvent des agrégations de données, comme le nombre total d'accidents survenus dans une région. La manière de définir les liens entre différentes régions est ce qu'on appelle le voisinage. Ces agrégations se basent fréquemment sur des frontières administratives, qui sont généralement arbitraires. Cependant, lorsque la division de la région ne correspond pas au processus spatial sous-jacent, ce qui est fréquent en raison du caractère arbitraire de ces frontières, cela peut entraîner une autocorrélation spatiale. Dans ce projet, les données québécoises dont nous disposons utilisent la municipalité comme niveau d'information spatiale le plus fin. L'analyse de données surfaciques est donc essentielle pour tirer des conclusions valides pour l'ensemble de la province. Dans cette section, nous définirons d'abord différents types de voisinages, puis nous examinerons les mesures d'autocorrélation spatiale, et enfin, nous aborderons des méthodes de modélisation des données surfaciques en fonction des liens de voisinage.

2.5.1 Voisinage

Une fois les données agrégées par zone géographique, il est nécessaire de capturer la proximité spatiale entre les régions. Pour ce faire, on considère que le processus spatial se propage à travers des relations de voisinage : un graphe de voisinage relie ainsi les sous-régions entre elles. Ces liens peuvent être orientés et pondérés, bien que l'on préfère souvent des liens non orientés et non pondérés pour garantir la symétrie : si i est voisin de j, alors j est également voisin de i (E. PEBESMA et BIVAND, 2023). On n'ajuste généralement pas pour les effets de bords, même si une sous-région proche du bord de la région étudiée aura moins de voisins.

Il existe plusieurs types de voisinages différents, mais ils n'assurent pas tous la sy-

métrie des relations, bien que cette symétrie puisse souvent être forcée. De plus, il est souvent préférable d'avoir un voisinage où toutes les sous-régions sont incluses. Parfois, des contraintes géographiques, comme des îles, peuvent isoler certaines sous-régions et former des voisinages indépendants du reste. En particulier, les processus de Markov nécessitent un graphique de voisinage connecté (un seul grand réseau) et non-dirigé (E. PEBESMA et BIVAND, 2023). Nous définirons plus tard les processus de Markov, mais notons qu'il existe certains ajustements permettant d'utiliser des voisinages non-connectés. La figure 2.15 montre différents types de voisinage. Le centroïde des régions a été utilisé pour les représenter. Examinons d'abord ces types de voisinages.

Voisinage contigu Il s'agit sûrement de la technique la plus intuitive : une région qui en touche une autre est considéré comme étant un voisin de cette région. Bien entendu, le graphique de voisinage engendré sera symétrique. Le seul choix à faire est généralement si deux régions sont contiguës lorsqu'elles partagent seulement un point ou plusieurs points de leur frontière. Deux régions avec un seul point en commun sur leur frontière seront dites contiguës de type « tour » , et si elles partagent plus d'un point, contiguës de type « reine » (MORAGA, 2023). Ces appellations proviennent des échecs.

Voisinage d'ordre k selon la contiguïté Il est possible d'aller au-delà d'un seul voisin contigu et de définir des régions comme étant voisine si elles sont à k régions contiguës l'une de l'autre. Encore une fois, il est possible que ces contiguïtés soit de type « tour » ou « reine ».

Voisinage basé sur les K plus proches voisins Comme son nom l'indique, les k plus proches voisins d'une région seront ses voisins. C'est-à-dire que même si deux régions se touchent, elles ne sont pas nécessairement voisines si l'une d'elles est plus proche de deux autres régions. Pour définir si une région est plus proche d'une autre, il faut définir un point représentant l'ensemble de la région. Le centroïde est souvent utilisé, mais on pourrait aussi utiliser d'autres points, comme par exemple le point où la densité

de population est la plus forte dans cette région. Le graphique de voisinage produit ne sera pas symétrique (MORAGA, 2023) mais il est possible de forcer la symétrie en ajoutant des liens.

Voisinage basé sur la distance Il est aussi possible de définir des régions comme étant voisines si elles se trouvent à une certaine distance l'une de l'autre. La distance entre deux régions est calculée entre deux points. On prend généralement les centroïdes comme point représentant la région. En prenant une distance assez grande, il est possible d'en trouver une telle qu'il n'y aura pas de régions sans voisins. De plus, le voisinage sera forcément symétrique. Il faut cependant être conscient que des plus petites régions auront beaucoup plus de voisins que des grandes régions. Ceci peut parfois être censé puisque les plus petites régions administratives sont généralement là où il y a plus de population. Ce n'est pas le cas avec les municipalités du sud du Québec, où les grandes municipalités sont généralement celles avec le plus de population. Dans le nord par contre, les municipalités ont tendance a être énorme avec peu d'habitants. Un voisinage basé sur la distances n'est sûrement pas idéal pour les municipalités du Québec pour étudier le nombre d'accidents par municipalités.

Voisinage basé sur une triangulation de Delaunay Des voisins par triangulation de Delaunay forment une triangulation de Delaunay où les sommets/nœuds de la triangulation sont les points représentant une région. Les liens de voisinage sont définis par la triangulation et s'étendent jusqu'à l'enveloppe convexe des points, c'est-à-dire le plus petit polygone convexe qui englobe l'ensemble des points (BIVAND, E. J. PEBESMA et GOMEZ-RUBIO, 2008b). Par construction, un voisinage basé sur une triangulation de Delaunay est toujours symétrique. Ce type de graphe de voisinage ressemble à celui d'un voisinage contigu, mais il inclut des liens pouvant atteindre des distances plus grandes, en particulier ceux connectant les points situés sur l'enveloppe convexe.

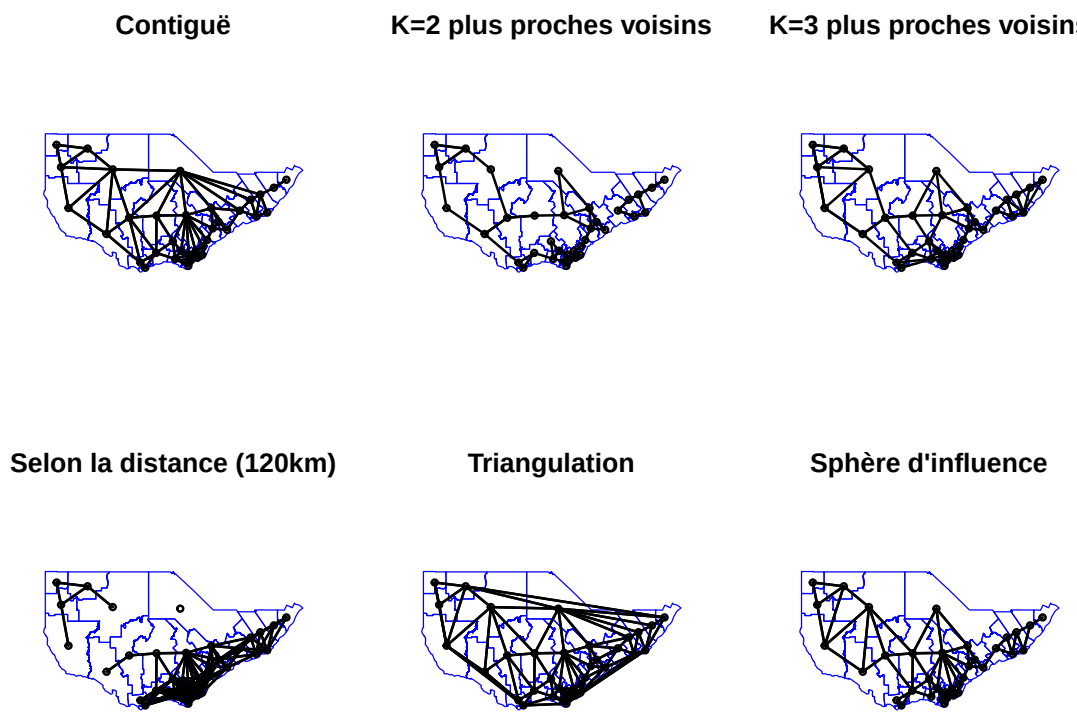

FIGURE 2.15 – Différents liens de voisinage entre MRC des régions administratives québécoises suivantes : Capitale-Nationale, Mauricie, Lanaudière, Laurentides, Outaouais et Abitibi-Témiscamingue.

Voisinage basé sur une sphère d'influence Le graphe par sphère d'influence est dérivé de la triangulation de Delaunay où on supprime les liens qui sont anormalement longs, souvent ceux formant l'enveloppe convexe. Supposons que nous ayons les points A et B. Pour chacun de ces points, trouvons le point le plus proche, disons qu'il s'agit respectivement des points C et D. A et B seront considérés comme voisins par sphère d'influence si des disques centrés sur ces deux points, dont les rayons sont égaux à la distance au voisin le plus proche (soit $|A-C|$ et $|B-D|$ respectivement), se croisent en deux points. Si ces disques se croisent en deux points, alors A et B sont voisins par sphère d'influence.

Matrice de voisinage Maintenant que nous avons établi les liens entre voisins, il est possible de définir une matrice représentant la structure de voisinage $\boldsymbol{W}$. Voici quelques éléments définissant la structure de $\boldsymbol{W}$:

1. $\boldsymbol{W}$ est une matrice carrée recouvrant les n sous-régions de la région étudiée.
2. w_{ij} représente le lien entre la sous-région i et la sous-région j (où i et $j = 1, 2, \ldots, n$) :
   a) w_{ij} peut être binaire, égal à 1 si i et j sont voisins et 0 sinon,
   b) w_{ij} peut aussi être l'inverse de la distance entre i et j,
   c) w_{ij} peut aussi être normalisé par rangée. On normalise généralement lors du calcul de l'autocorrélation. Tous les liens sortant de i auront un poids de $(\text{nombre de voisins}_i)^{-1}$,
   d) w_{ii} vaut 0.

Notons que les liens binaires et inverses de la distance conservent la symétrie du voisinage. Par contre, les liens normalisés par rangée ne conservent pas la symétrie, car les poids des liens entre deux voisins peuvent être différents dans chaque direction.

La figure 2.16 montre la matrice de voisinage contiguë des régions de la figure 2.15 avec des liens binaires et pondérées par l'inverse de la distance. Des poids plus élevés dans la matrice de voisinage par inverse de distance indiquent une plus grande proximité entre les régions. On observe que la matrice est symétrique, mais le fait que les valeurs suivent la diagonale ne signifie rien puisque l'ordre des zones est arbitraire.

2.5.2 Autocorrélation spatiale

Contrairement à la corrélation, qui évalue le lien linéaire entre deux variables distinctes, l'autocorrélation spatiale examine le lien spatial entre les valeurs d'une même variable mesurées à différents endroits. Elle est dite positive lorsque des observations ayant des valeurs similaires sont proches spatialement, et négative lorsque des observations de valeurs dissemblables sont regroupées spatialement (MORAGA, 2023). En d'autres termes, en présence d'autocorrélation spatiale, la valeur observée en un point peut être prédite en fonction de la valeur d'un point proche. Cette autocorrélation peut résulter de

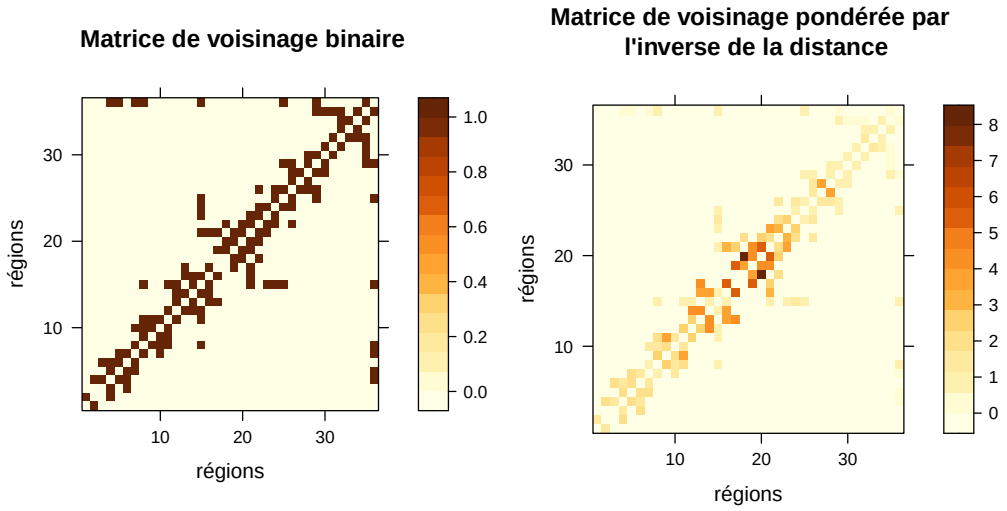

FIGURE 2.16 – Matrice de voisinage contiguë avec pondération différente.

plusieurs facteurs : des frontières administratives qui ne reflètent pas le processus spatial sous-jacent, l'omission de variables pertinentes dans le modèle, ou la corrélation avec une autre variable spatialement autocorrélée (par exemple, la densité de population ou le réseau routier, qui influencent respectivement le trafic ou les accidents). L'autocorrélation spatiale devient ainsi un indicateur précieux pour détecter une spécification inadéquate du modèle. Il existe plusieurs mesures d'autocorrélation spatiale. Cependant, ces mesures partagent une limitation : elles peuvent interpréter des erreurs dans la spécification du modèle comme de l'autocorrélation (E. PEBESMA et BIVAND, 2023).

Mesure globale : Indice *I* de Moran

Cette mesure considère le niveau moyen de l'autocorrélation spatiale à travers l'ensemble des observations de la variable étudiée. Il en existe d'autres que celles présentées ici comme l'indice *C* global de Geary ou celui de Getis–Ord.

L'indice *I* de Moran (MORAN, 1950) peut être calculé de la manière suivante :

$$I = \frac{n \sum_i \sum_j w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \tilde{Y})^2}, \tag{2.26}$$

où n est le nombre de zones, Y_i, l'observation de la variable dans la région i, $\bar{Y}$, la moyenne des observations, w_{ij} le poids spatial accordé au lien entre deux régions pour $i, j = 0, 1, 2, \ldots, n$. L'indice I se situe généralement dans un intervalle de -1 à 1. Généralement, on veut tester l'hypothèse nulle H_0 selon laquelle il n'y a pas d'autocorrélation spatiale. Sous H_0, les observations Y_i sont indépendantes et identiquement distribuées et I est asymptotiquement normal de moyenne $\mathrm{E}[I]$ et variance Var(I) connues :

$$\mathrm{E}[I] = \frac{-1}{n-1}, \ \mathrm{Var(I)} = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}, \tag{2.27}$$

où $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum_{i \neq j}(w_{ij} + w_{ji})^2$ et $S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik}\right)^2$. Il suffit de calculer la statistique de test z

$$z = \frac{I - \mathrm{E}[I]}{\sqrt{\mathrm{Var(I)}}}. \tag{2.28}$$

Sous H_0, z suit une distribution normale centrée réduite (MORAGA, 2023).

Un indice I significativement plus petit que $E[I]$ implique une autocorrélation spatiale négative, alors qu'un I significativement plus grand que $E[I]$ implique une autocorrélation spatiale positive. Finalement, un indice I autour de $E[I]$ implique un caractère spatial aléatoire. Par conséquent, il est possible de tester H_0 contre soit H_1, il y a autocorrélation spatiale, soit H_1, il y a autocorrélation spatiale positive, ou bien, H_1, il y a autocorrélation spatiale négative. L'hypothèse d'une autocorrélation spatiale positive est généralement l'hypothèse alternative par défaut dans les implémentations.

Il existe une autre manière de calculer la variance de I. Cette technique utilise une approche Monte-Carlo. De plus, on n'a plus à supposer quoi que ce soit sur la distribution de I. D'abord, on simule m jeux de données où l'on permute les valeurs entre les régions et on calcule l'indice I pour chacune des permutations. On obtient donc m indices I avec les données permutées. Il ne reste plus qu'à comparer l'indice I des vraies observations à cette distribution de I. Le résultat du test de Monte-Carlo est souvent très semblable à celui où sous H_0, $z \sim \mathcal{N}(0,1)$.

Il existe également une version bayésienne de l'indice de Moran. L'indice de Moran de Bayes empirique est utilisé pour tester l'autocorrélation spatiale des taux, comme

mentionné par ASSUNCÃO et REIS, 1999. On a donc besoin d'un vecteur d'une variable de comptage et d'un vecteur d'exposition. En épidémiologie, l'exposition est souvent la population à risque (BIVAND et WONG, 2018).

Pour conclure, les mesures globales sont plutôt rudimentaires et sont victimes de toutes sortes d'erreurs de spécification, pas seulement celle d'autocorrélation spatiale. Cependant, les mesures locales offrent une approche complémentaire en se concentrant sur les variations spécifiques à chaque région.

Mesures locales d'autocorrélation spatiale

En réponse aux limitations des mesures globales, ANSELIN, 1995 introduit les indicateurs locaux d'association spatiale (LISA pour *Local Indicators of Spatial Association*). Ces mesures permettent de détecter les regroupements spatiaux significatifs de valeurs similaires autour de chaque observation. La somme des valeurs LISA pour toutes les observations est idéalement un multiple d'un indicateur global. Une bonne porte d'entrée vers les LISA est le diagramme de Moran.

Le diagramme de Moran est un nuage de points avec les valeurs de la variable d'intérêt sur l'axe des x et la moyenne lissée spatialement sur l'axe des y. La moyenne lissée spatialement est la moyenne pondérée de la valeur des voisins, autrement dit, la valeur estimée à partir des voisins. La figure 2.17 montre un exemple de diagramme de Moran où nous évaluons le revenu médian des ménages par MRC dans le sud du Québec, en utilisant un voisinage contigu et des liens normalisés par rangées. Les points en rouge dans le diagramme de Moran représentent les observations très éloignées des autres. Ces points sont qualifiés de points à valeur extrême. Il est plus fréquent d'utiliser un voisinage avec des liens pondérés par normalisation des rangées, ce qui permet de mettre les deux axes sur une même échelle et facilite la comparaison. On peut diviser le diagramme de Moran en quadrants : le quadrant en bas à gauche est qualifié de *low-low*, indiquant un regroupement de points dont les valeurs observées et prédites sont inférieures à la moyenne. Le quadrant en haut à droite est qualifié de *high-high*, indique un regroupement de points dont les valeurs observées et prédites sont supérieures à la moyenne. Les points

dans ces deux quadrants montrent une autocorrélation spatiale positive. Le quadrant en haut à gauche est appelé *low-high*, et enfin, le quadrant en bas à droite est appelé *high-low*. Des points dans ces deux quadrants montrent une autocorrélation spatiale négative. Les quadrants sont délimités par les moyennes des valeurs de la variable étudiée et de ses valeurs lissées spatialement. On s'attend généralement à ce que la majorité des points se trouvent le long de la diagonale dans les quadrants *low-low* et *high-high*, ce qui reflète une autocorrélation spatiale positive.

Il est possible de calculer le *hat-value* pour chaque point du diagramme de Moran. Un *hat-value* élevée (proche de 1) indique une valeur extrême sur l'axe des x, c'est-à-dire une valeur extrême de la variable étudiée. Cette extrême sur l'axe des x suggère que l'observation exerce une influence importante sur les résultats du modèle, car la valeur de la région i diffère de manière significative de la valeur prédite par ses voisins. En d'autres termes, un *hat-value* élevée peut indiquer que l'observation est un point ayant un impact disproportionné par rapport aux autres observations dans un modèle (BERGMANN, 2015). Voici comment un *hat-value* est calculé :

$$h_i = \frac{1}{n} + \frac{(Y_i - \bar{Y})^2}{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}, \tag{2.29}$$

où les Y_i sont les valeurs des zones, et n, le nombre de zones.

On remarque que les points à valeurs extrêmes et les points à *hat-value*s élevés sont semblables. La seule valeur extrême qui n'a pas un *hat-value* élevé provient du voisin d'un *hat-value* élevé, et est sûrement extrême sur l'axe des y simplement parce que son voisin est extrême sur l'axe des x.

Il existe encore d'autres mesures locales comme le C de Geary et le G de Getis-Ord, mais nous ne présenterons que l'indice I de Moran local. Notons que les indices de Geary et Getis-Ord peuvent être utilisés à la fois de manière locale ou globale, comme celui de Moran. L'indice I de Moran local est le LISA le plus commun. On peut calculer le I de Moran pour l'observation i comme ceci (MORAGA, 2023) :

$$I_i = \frac{n(Y_i - \bar{Y})}{\sum_j (Y_j - \bar{Y})^2} \sum_j w_{ij}(Y_j - \bar{Y}), \tag{2.30}$$

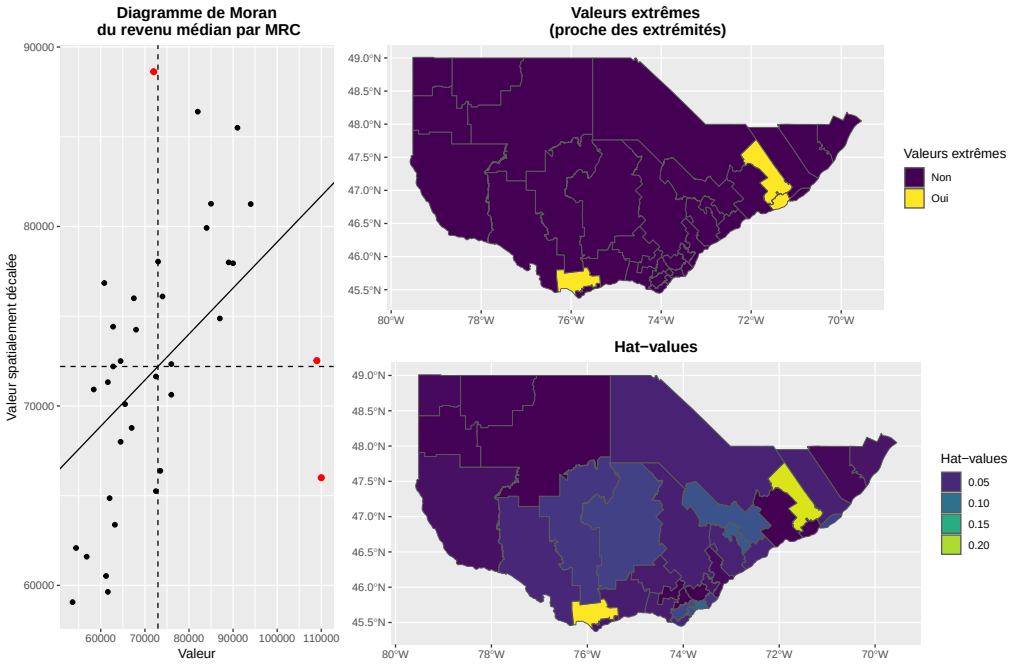

FIGURE 2.17 – À gauche, le diagramme de Moran. Les points en rouge sont les valeurs extrêmes. Ces points sont représentés sur la carte en haut à droite. Les *hat-value*s sont sur la carte en bas à droite.

où n est ici aussi le nombre total de régions, Y_i, l'observation de la région i.

On peut retrouver l'indice I de Moran global à partir des indices locaux :

$$I = \frac{\sum_i I_i}{\sum_{i \neq j} w_{ij}}. \tag{2.31}$$

Chaque indice local I_i peut être testé statistiquement à l'aide d'un test z. On examine alors si les valeurs observées dans la région i et celles des régions voisines diffèrent significativement de ce qui serait attendu sous l'hypothèse nulle d'absence d'autocorrélation spatiale. En absence d'autocorrélation spatiale locale, la valeur de la région i serait indépendante de celles de ses voisines, et les écarts observés entre régions seraient uniquement dus au hasard. En supposant la normalité des I_i, l'espérance et la variance sont connues et détaillées dans ANSELIN, 1995. On obtient donc des valeurs p pour chaque observation.

Cependant, comme les mesures locales sont calculées pour chaque région, le nombre de tests statistiques à effectuer équivaut au nombre de régions, ce qui pose un problème de comparaisons multiples. C et SINGER, 2006 conclut donc qu'il est préférable d'ajuster

pour le taux de fausses découvertes. Le taux de fausse découverte est le taux d'évènements jugés significatifs alors qu'ils ne le sont réellement pas. Des exemples d'ajustement de taux de fausses découvertes sont ceux de Bonferroni et ceux présentés dans BENJAMINI et YEKUTIELI, 2001 et BENJAMINI et HOCHBERG, 1995.

2.5.3 Modélisation de données surfaciques

La littérature sur les modèles de données surfaciques est très vaste, particulièrement dans le domaine de l'épidémiologie. En effet, pour des raisons de confidentialité, il est souvent impossible d'obtenir des données géolocalisées des individus malades. On se contente alors d'agrégations par région, c'est-à-dire de données surfaciques. En épidémiologie, on étudie la distribution spatiale des maladies et on identifie les régions avec des risques relatifs plus ou moins élevés (BLANGIARDO et CAMELETTI, 2015).

Bien entendu, on pourrait simplement estimer les risques relatifs avec des ratios d'incidence standardisées. Cependant, il est préférable de modéliser les risques relatifs avec des modèles bayésiens hiérarchiques qui permettent d'utiliser l'information provenant du voisinage de la région, et qui gèrent beaucoup mieux des régions avec peu de population ou de faibles comptes (P. MORAGA, 2019). De plus, les modèles bayésiens hiérarchiques permettent de facilement sélectionner des modèles à l'aide du DIC. On rappelle qu'un plus petit DIC indique un modèle qui s'ajuste mieux aux données.

Modèles bayésiens hiérarchiques

Nous allons présenter ici trois modèles bayésiens hiérarchiques permettant de prendre en compte l'information du voisinage : le modèle autorégressif conditionnel (CAR), le modèle autorégressif conditionnel intrinsèque (ICAR) et le modèle Besag–York–Mollié (BYM).

Les modèles sont spécifiés dans le cadre des modèles linéaires généralisés où Y_i représente la valeur de la variable réponse pour la région i, où $i = 1, \ldots, n$. On note aussi $X_i = (X_{i1}, \ldots, X_{ip})$ le vecteur de p covariables mesurées pour la région i. Les modèles que

nous présentons sont des modèles à effets mixtes, qui prennent la forme de base suivante :

$$\ln\left(\mathrm{E}[Y_i|X_i,b_i]\right) = X_i\boldsymbol{\beta} + b_i,\ i = 1,\ldots,n \tag{2.32}$$

où $\boldsymbol{\beta}$ représente le vecteur des coefficients (effets fixes) du modèle et b_i représente un effet aléatoire spatial correspondant à la zone i tel que

$$\boldsymbol{b} = (b_1,\ldots,b_n) \sim \mathscr{MN}(0,\sigma_b^2\Sigma),$$

où $\sigma_b^2\Sigma$ représente la matrice $n \times n$ de variance-covariance des effets aléatoires spatiaux et $\mathscr{MN}$ désigne une distribution normale multivariée.

Les modèles CAR et ICAR modélisent uniquement la structure spatiale à travers b_i. En revanche, le modèle BYM inclut aussi une composante supplémentaire : un effet aléatoire non structuré $v_i \sim \mathscr{MN}(0,\sigma_v^2)$, permettant de capturer la variabilité indépendante de la structure spatiale.

Dans ces trois modèles, la structure de voisinage est modélisée à l'aide d'un champ gaussien de Markov, qui est une généralisation spatiale du processus de Markov. Dans ce contexte, la dépendance spatiale est capturée à travers une structure conditionnelle définie par les relations de voisinage, où chaque région i est influencée uniquement par ses voisines immédiates. Un processus de Markov est un processus stochastique temporel qui a la propriété de Markov. Cette propriété précise que l'effet aléatoire b_i est indépendant de tous les autres paramètres, notés $\boldsymbol{b}_{-i}$, sachant l'ensemble de ses voisins $N(i)$ (BLANGIARDO et CAMELETTI, 2015) :

$$b_i \perp\!\!\!\perp \boldsymbol{b}_{-i} \mid \boldsymbol{b}_{N(i)}. \tag{2.33}$$

Par conséquent, on a aussi que pour tout i et j non-voisins,

$$b_i \perp\!\!\!\perp b_j \mid \boldsymbol{b}_{-ij}, \tag{2.34}$$

où $\boldsymbol{b}_{-ij}$ représente l'ensemble des paramètres $\boldsymbol{b}$ sauf les paramètres i et j. La propriété Markovienne a l'avantage d'engendrer des matrices de précision $\boldsymbol{Q} = \Sigma^{-1}$ très clairsemées, car

$$Q_{ij} = 0, \text{ pour tout } i,j \text{ non-voisins.}$$

Il y a donc d'importants bénéfices computationnels.

La structure des effets aléatoires $\boldsymbol{b}$ est d'un intérêt particulier, et donc définir $\boldsymbol{Q} = \Sigma^{-1}$, la matrice de précision, est essentiel pour décrire les dépendances spatiales. C'est d'ailleurs la structure de $\boldsymbol{b}$ qui varie dans les modèles zonaux courants. En effet, les modèles autorégressifs conditionnels, les modèles autorégressifs conditionnels intrinsèques, ainsi que leurs extensions, comme les modèles Besag–York–Mollié (BYM) et Leroux (LEROUX, LEI et BRESLOW, 2000) se distinguent principalement par la structure de $\boldsymbol{b}$. Dans ces modèles, la dépendance spatiale entre les effets aléatoires et leurs voisins est spécifiée de manière explicite via une distribution conditionnelle, faisant de ces modèles des cas particuliers de champs de Markov.

Modèles autorégressifs conditionnels (CAR) BESAG, 1974 présente le modèle CAR. En considérant n régions, ayant chacune un ensemble de voisins $N(i)$, la distribution conditionnelle des effets aléatoires b_i est alors (BLANGIARDO et CAMELETTI, 2015) :

$$b_i \mid \boldsymbol{b}_{-i} \sim \mathcal{N}\left(\mu_i + \frac{\phi}{N_i}\sum_{j=1}^{n} w_{ij}(b_j - \mu_j), s_i^2\right), \tag{2.35}$$

où μ_i est la moyenne de la région i, w_{ij} est le lien de voisinage binaire entre i et j ($w_{ii} = 0$), le paramètre ϕ contrôle l'effet de la proximité spatiale et N_i est le nombre de voisins de la région i ($N_i = \#N(i)$). Finalement, on a $s_i^2 = \sigma_b^2/N_i$. σ_b^2 est la variance des effets aléatoires structurés spatialement. Une région avec plusieurs voisins (grand N_i) aura donc une plus petite variance et ceci est logique puisque plus de voisins signifie plus d'information sur la valeur de l'effet aléatoire b_i.

La matrice de voisinage $\boldsymbol{W}$ est définie par $W_{ij} = w_{ij}/N_i$. La matrice de précision $\boldsymbol{Q}$ des effets aléatoires spatiaux est alors définie par $\boldsymbol{Q} = \boldsymbol{I} - \phi\boldsymbol{W}$ (E. PEBESMA et BIVAND, 2023). Cependant, cette spécification est rarement utilisée car le paramètre ϕ est difficile à estimer (BLANGIARDO et CAMELETTI, 2015).

Modèles autorégressifs conditionnels intrinsèques (ICAR) Il s'agit d'une version simplifiée du modèle CAR, où ϕ est fixé à 1. Dans ce cas, la distribution conditionnelle des

b_i devient :

$$b_i \mid \boldsymbol{b}_{-i} \sim \mathcal{N}\left(\mu_i + \frac{1}{N_i}\sum_{j=1}^{n} w_{ij}(b_j - \mu_j), s_i^2\right),$$

et la matrice de précision devient :

$$\boldsymbol{Q} = \text{diag}(n_i) - \boldsymbol{W},$$

où les n_i sont les sommes des rangées de $\boldsymbol{W}$ (E. PEBESMA et BIVAND, 2023). Cependant, en fixant ϕ à 1, la matrice de covariance n'est plus définie positive. Cela peut résulter en une dépendance parfaite entre certaines régions voisines, rendant la matrice singulière et la distribution conjointe des effets aléatoires b non définie (BLANGIARDO et CAMELETTI, 2015). C'est pour résoudre ce problème que le modèle BYM a été développé.

Modèle Besag–York–Mollié (BYM) Il s'agit d'un des modèles les plus courants pour la modélisation spatiale (MORAGA, 2023). Il a été développé par BESAG, YORK et MOLLIÉ, 1991. C'est une extension d'un modèle ICAR, où on fixe μ_j à 0, et on se retrouve avec la distribution conditionnelle des b_i suivante :

$$b_i \mid \boldsymbol{b}_{-i} \sim \mathcal{N}\left(\frac{1}{N_i}\sum_{j=1}^{n} w_{ij} b_j, s_i^2\right).$$

De plus, on ajoute un effet aléatoire échangeable non-structuré qui modélise le bruit non correlé (P. MORAGA, 2019). Ce nouvel effet aléatoire v_i suit une distribution $\mathcal{N}(0, \sigma_v^2)$.

$$Y_i \mid X_i, b_i, v_i \sim \mathcal{N}(\mu_i, \sigma^2), i = 1, \ldots, n,$$

$$\mu_i = X_i\beta + b_i + v_i.$$

Nous avons décrit jusqu'à maintenant les modèles dans leur forme générale (normale). Dans notre cas, nous modélisons des décomptes d'accidents par région. Ce genre de variables est généralement modélisé en utilisant une distribution de Poisson de paramètre $\lambda_i = E_i\theta_i$. Ici $\theta_i = e^{\eta_i}$ représente le risque relatif de la région i. Le prédicteur linéaire $\eta_i = X_i\beta + b_i + v_i$ inclut les variables explicatives et les effets aléatoires. E_i est une mesure de l'exposition de la région i et représente le nombre attendus de cas. En épidémiologie,

il est courant de normaliser le nombre total d'observations par un taux de population. On peut représenter le nombre attendu de cas de la manière suivante :

$$E_i = r \cdot n_i,$$

où

$$r = \frac{\text{nombre total d'observations}}{\text{population totale}},$$

et

$$n_i = \text{population dans la région } i.$$

On se retrouve donc avec

$$E_i = \frac{\text{population dans la région } i}{\text{population totale}} \cdot \text{nombre total d'observations.}$$

Ce taux de normalisation basé sur la population est bien utile lorsque l'exposition est mesurée en termes de population, comme en épidémiologie. Ce n'est cependant pas le cas pour les accidents de la route. On préfère alors utiliser comme « population » la quantité de déplacement totale, dans notre cas, le VKT, définis à la section 1.1.1. Contrairement à la population, le VKT reflète directement la quantité de déplacements, rendant cette mesure d'exposition plus appropriée pour la sécurité routière. Les mêmes formules peuvent alors être utilisées pour obtenir les nombres attendus d'accidents par région, E_i, en remplaçant les mesures de population par des mesures de VKT, et le nombre total d'observations par le nombre total d'accidents.

Il peut également être intéressant de calculer la proportion de variance expliquée par la structure spatiale, tel que proposé dans BLANGIARDO et CAMELETTI, 2015. En effet, les effets aléatoires b_i et v_i du modèle BYM ont chacun une variance. La variance de l'effet aléatoire avec structure spatiale est notée σ_b^2, tandis que la variance de l'effet aléatoire sans structure spatiale est notée σ_v^2. En ajoutant l'option *scale.model=T* lors de la modélisation sur *INLA*, les variances sont sur une échelle comparable. On calcule facilement la proportion expliquée par la structure spatiale une fois le modèle ajusté de la manière suivante :

$$\text{Proportion de la variance expliquée par la structure spatiale} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_v^2}. \tag{2.36}$$

Il reste encore un problème dans les cas où trop de décomptes d'accidents sont nuls (c'est-à-dire dans le cas d'un excès de zéros), d'autres distribution que celle de Poisson sont plus adaptées, comme une binomiale-négative ou une Poisson zéro-enflée. Ces deux distributions traitent l'excès de zéros différemment.

La distribution binomiale-négative permet à la variance d'être plus grande que la moyenne, alors qu'elle est égale à la moyenne dans une distribution de Poisson. Sans modéliser directement l'excès de zéros, en permettant cette surdispersion, la distribution binomiale-négative peut indirectement le modéliser.

La distribution *zero-inflated* Poisson quant à elle modélise directement l'excès de zéros. Cette distribution est en fait un mélange de deux distributions, une Poisson et une Bernoulli. La distribution de Bernoulli traite les excès de zéros, en mettant une probabilité sur les zéros qui sont issus d'un mécanisme différent du processus sous-jacent (zéros structurellement non liés aux données) (BLANGIARDO et CAMELETTI, 2015). Dans le cas des accidents agrégés par zone, un zéro structurel correspond à une zone où il est impossible qu'un accident se produise (par exemple, une zone non habitée ou sans routes).

Extensions aux modèles spatio-temporels

Comme nous avons des données sur les accidents entre 2011 et 2022, il devient pertinent d'intégrer une composante temporelle en plus de la composante spatiale. Encore une fois, les modèles spatio-temporels proviennent de la littérature en épidémiologie. Nous sommes dans un contexte très similaire à celui des modèles spatiaux. Nous nous limiterons ici à présenter deux modèles spatio-temporels dans le cadre d'une distribution de Poisson :

$$Y_{ij} \sim \text{Poi}(E_{ij}\theta_{ij}),\ i = 1,\ldots,I,\ j = 1,\ldots,J$$

où i est une région et j un instant.

Le premier modèle est celui de Bernardinelli (BERNARDINELLI et al., 1995), un modèle commun en épidémiologie :

$$\ln(\theta_{ij}) = X_{ij}\beta + b_i + \nu_i + (\omega + \delta_i) * t_j, \tag{2.37}$$

On retrouve en quelque sorte le modèle BYM avec les effets aléatoires b_i et ν_i qui ont respectivement des structures ICAR (centrées à zéro) et iid. La nouveauté vient des termes $\omega \times t_j$ et $\delta_i \times t_j$, où t_j est simplement le temps. Le paramètre ω peut alors être vu comme une tendance globale temporelle linéaire, et δ_i, appelé tendance différentielle, comme une mesure qui indique à quel point la tendance temporelle de la région i diffère de la tendance globale ω. θ_{ij} représente ici aussi le risque relatif.

Le second modèle est celui de Knorr-Held. Le modèle de Bernardinelli suppose une tendance temporelle linéaire. Dans le cas où cette supposition n'est pas réaliste, le modèle présenté par (KNORR-HELD, 2000) modélise plutôt l'effet temporel avec des effets aléatoires :

$$\ln(\theta_{ij}) = X_{ij}\beta + b_i + \nu_i + \gamma_j + \phi_j + \delta_{ij}. \tag{2.38}$$

Ce modèle conserve la structure BYM pour les effets spatiaux (b_i et ν_i). En revanche, pour les effets aléatoires temporels, deux composantes distinctes sont introduites.

— γ_j est un effet temporel structuré qui peut suivre, par exemple, une marche aléatoire du premier ou du second ordre ;

— ϕ_j est un effet temporel non structuré (iid).

Une marche aléatoire est un processus stochastique où chaque valeur dépend conditionnellement des valeurs précédentes. On distingue notamment :

— Une marche aléatoire du premier ordre :

$$\gamma_j \mid \gamma_{j-1} \sim \mathcal{N}(\gamma_{j-1}, \sigma_\gamma^2),$$

— Une marche aléatoire du second ordre :

$$\gamma_j \mid \gamma_{j-1}, \gamma_{j-2} \sim \mathcal{N}(2\gamma_{j-1} - \gamma_{j-2}, \sigma_\gamma^2).$$

Enfin, l'effet aléatoire δ_{ij} représente une interaction entre les dimensions spatiale et temporelle. Ces interactions, proposées par KNORR-HELD, 2000, combinent des effets structurés (b_i ou ν_i) et des effets temporels (ϕ_j ou γ_j). Quatre spécifications sont possibles :

— b_i (structuré spatial), ϕ_j (non structuré temporel) ;

— b_i (structuré spatial), γ_j (structuré temporel) ;

— v_i (non structuré spatial) , ϕ_j (non structuré temporel) ;

— v_i (non structuré spatial), γ_j (structuré temporel) .

Ces interactions permettent de modéliser plus finement les relations complexes entre les effets spatiaux et temporels.

2.5.4 Problèmes des données surfaciques

Dans la section sur l'autocorrélation spatiale, nous avons mentionné que le découpage d'une région en sous-régions pouvait engendrer une autocorrélation spatiale. Il existe d'autres problèmes avec les données surfaciques.

Problème des unités spatiales modifiables

Connu en anglais comme le *Modifiable Areal Unit Problem* (MAUP), le problème des unités spatiales modifiables fait référence au fait que les résultats d'une analyse spatiale peuvent changer en fonction de l'agrégation de ces mêmes données. Par exemple, les résultats peuvent être différents si l'on agrège les résultats du nombre d'accidents par municipalités ou bien par MRC.

Problème des données non-alignées

Connu en anglais sous le nom de *Misaligned Data Problem* (MIDP) (BANERJEE et al., 2003), le problème des données non-alignées se produit lorsque des données qui n'ont pas été recueillies au même niveau sont utilisées ensemble dans une même analyse. Par exemple, en prenant des données provenant de Statistique Canada, les municipalités canadiennes sont délimitées par leur limite terrestre, alors que les limites administratives des municipalités incluent les cours d'eau et recouvrent tout le territoire. Les données provenant de ces deux sources peuvent être différentes, comme l'aire de la municipalité.

Finalement, la modélisation de données surfaciques utilise des modèles plus simples à implémenter que la modélisation de MPS. C'est notamment parce qu'ils ont été développés il y a plus longtemps. Cette modélisation peut cependant devenir compliquée simplement en raison du grand nombre de modèles possibles. Dans cette recherche, nous tenterons de garder des modèles relativement simples puisque leur interprétation sera plus facile.

Chapitre 3

Exploration et préparation des données

Maintenant que nous avons présenté plusieurs méthodes d'analyse des données spatiales dans la revue de littérature, ce chapitre vise à introduire les données sur lesquelles ces analyses seront appliquées. La première partie de ce chapitre se concentre sur les données relatives aux accidents, qui proviennent de la Société de l'assurance automobile du Québec (SAAQ), mais sont collectées par des policiers. Bien que des données similaires soient accessibles publiquement, celles que nous avons reçues contiennent des informations plus détaillées. Le niveau spatial le plus fin disponible dans ces données correspond à la municipalité où l'accident s'est produit. La Ville de Montréal, en s'appuyant sur les données de la SAAQ, produit également des données ponctuelles sur les accidents sur l'île de Montréal. La deuxième section de ce chapitre abordera les données externes que nous avons collectées pour enrichir l'analyse. Enfin, ce chapitre se conclura par une présentation des étapes de prétraitement qui ont été appliquées à ces données.

Avant de poursuivre, notons que nous allons convertir tous les types de données au système canadien de référence spatiale, le NAD83, plutôt que d'utiliser le système géodésique mondial, le WGS84. Bien que ces deux systèmes présentent des différences minimes, ils restent très semblables. Nous avons choisi le NAD83 simplement puisqu'il s'agit du système géodésique canadien. La projection que nous utiliserons tout au long de ce mémoire est le système UTM, spécifiquement la zone 18, qui est celle incluant l'île de

Montréal. Ce choix de projection garantit une représentation précise des distances et des surfaces dans notre analyse régionale.

Nous avons également choisi le kilomètre comme unité de distance par défaut, plutôt que le mètre. Cette décision est motivée par le fait que l'utilisation de kilomètres favorise une meilleure convergence des modèles ponctuels, notamment lorsque le jeu de données est limité. En effet, cela permet d'éviter des échelles numériques trop petites dans nos calculs, ce qui peut améliorer la stabilité des résultats.

3.1 Données relatives aux accidents de la SAAQ

Les données relatives aux accidents fournies par la SAAQ comprennent des détails précis sur chaque conducteur impliqué dans un accident. Chaque ligne du jeu de données correspond à un conducteur impliqué. Ces informations, recueillies par les policiers du Québec, reflètent une évaluation subjective effectuée au moment de l'accident. Elles sont utiles pour orienter les actions de prévention en matière de sécurité routière. Les accidents sont comptabilisés dès qu'il y a des dommages corporels ou lorsque l'évaluation des dommages matériels est estimée à plus de 2000$.

Nous disposons de données sur tous les accidents comptabilisés du 1er janvier 2011 au 31 décembre 2022. La figure 3.1 présente le nombre d'accidents selon la gravité ainsi que le nombre de personnes blessées ou décédées par année.

Un accident est considéré comme mortel si une personne impliquée décède dans les 30 jours suivant l'accident. Un accident est qualifié de grave si aucune personne ne décède, mais qu'au moins une personne est gravement blessée (blessures nécessitant l'hospitalisation). Enfin, un accident est considéré comme léger s'il n'y a ni mort ni blessé grave, mais qu'au moins une personne est légèrement blessée.

En regardant maintenant la figure, on remarque que les tendances pour une même gravité sont très semblables. Par exemple, les tendances du nombre d'accidents mortels et du nombre de personnes décédées sont très similaires. Dans tous les cas, le total des accidents est simplement inférieur à celui des blessés ou des décès. Ceci est tout à fait

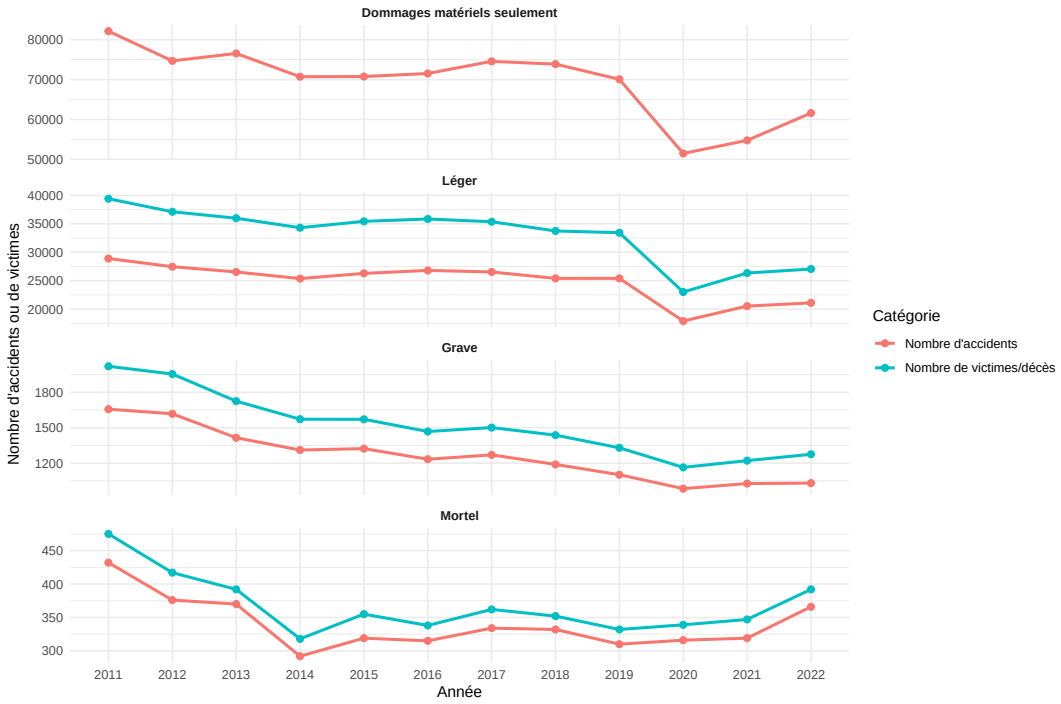

FIGURE 3.1 – Nombre d'accidents et nombre de blessés/décès annuels selon la gravité.

logique, puisqu'un accident grave peut causer plus d'un blessé grave. Dans la suite de ce mémoire, nous n'utiliserons que le nombre d'accidents de différente gravité de l'accident lui-même et pas le nombre de victimes ou décès.

En 2020, on observe une baisse notable du nombre d'accidents légers et des accidents impliquant uniquement des dommages matériels. Une diminution moins marquée, mais tout de même visible, est présente pour les accidents graves. En revanche, on observe une légère hausse pour les accidents mortels. Nous tenterons d'expliquer cette hausse dans ce mémoire.

Il semble y avoir eu un changement relativement important en 2014, où les accidents mortels atteignent un minimum avant de remonter, tandis que les accidents légers enregistrent un minimum local. Les accidents graves, quant à eux, ne semblent pas être affectés par cette année. Comme notre objectif principal est d'analyser la remontée des accidents et du nombre de victimes après le déclenchement de la pandémie de COVID-19, nous n'analyserons que les données à partir de 2014. De plus, nous retirerons dès main-

tenant les accidents impliquant uniquement des dommages matériels, car ceux-ci peuvent être sous-représentés ; en effet, tous les accidents avec seulement des dommages matériels ne sont pas nécessairement rapportés à la police. Ces accidents auraient pu servir d'indicateur de l'exposition, mais dans ce projet, nous avons préféré définir l'exposition par d'autres moyens, afin d'éviter une approche qui aurait pu être perçue comme circulaire.

Le jeu de données de la SAAQ fournit une quantité importante d'informations sur chaque accident. Sur les 126 variables disponibles, seules quatre variables concernent directement le conducteur ou le véhicule impliqué dans l'accident. Les 122 autres variables sont identiques pour les lignes correspondant à un même accident (mais bien entendu différente pour des accidents distincts). Les informations sur les accidents sont de nature spatiale, temporelle ou sur les caractéristiques de l'accident lui-même.

3.1.1 Données spatiales

Pour chaque accident, l'information spatiale la plus précise disponible est au niveau de la municipalité. Bien sûr, la municipalité régionale de comté (MRC) et la région administrative sont également accessibles. D'autres données spatiales sont disponibles, mais elles sont souvent éparpillées et présentent de nombreuses valeurs manquantes. Par exemple, les informations sur le numéro de la route, la rue où l'accident a eu lieu, et la borne kilométrique de la route sont également disponibles. Théoriquement, ces données permettraient de localiser les accidents avec une certaine précision. Il serait intéressant de développer un système de géocodage.

3.1.2 Données temporelles

Nous avons accès à la date de l'accident, ce qui nous permet de connaître le jour de la semaine, le mois et l'année. Nous avons également accès à l'heure de l'accident, mais seulement par intervalles de 60 minutes, par exemple : « 20h00m00s-20h59m59s » .

3.1.3 Données sur l'accident

Finalement, plusieurs données sur l'accident en tant que tel sont disponibles. Certaines de ces données devront être prétraitées et c'est ce que nous verrons dans cette section.

Vitesse autorisée

Tout d'abord, nous avons accès à la vitesse autorisée là où l'accident a eu lieu. La figure 3.2 montre les moyennes annuelles des vitesses permises dans les zones où se sont produits les accidents selon différentes gravités.

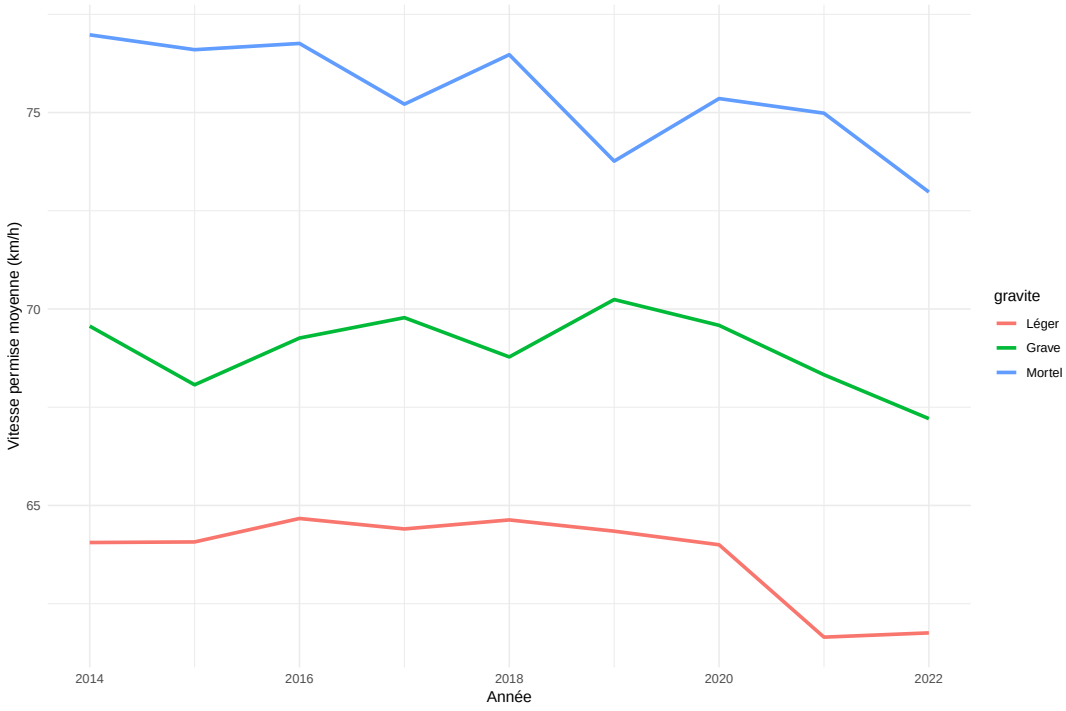

FIGURE 3.2 – Moyennes annuelles des vitesses permises sur les lieux où se sont produits les accidents légers, graves et mortels.

Comme on pouvait s'y attendre, plus la gravité des accidents augmente, plus ils se produisent dans des zones où la vitesse permise est élevée. De plus, depuis 2020, on observe que tous les types d'accidents ont lieu, en moyenne, dans des zones où la vitesse permise est plus faible qu'auparavant. Ceci n'est peut-être pas directement lié à la pandémie de COVID-19 et reflète peut-être simplement les réglementations mises en place

dans certaines grandes du Québec, où la vitesse permise a été réduite dans le cadre de politiques de sécurité routière visant à diminuer la gravité des accidents et à protéger les usagers vulnérables.

Genre d'accident

Le genre (ou type) de l'accident désigne le premier fait physique survenu lors de l'accident. Cette variable est codifiée en 32 catégories. Nous avons réduit ce nombre à 8 catégories en se basant sur le classement utilisé dans les données publiques québécoises (P. d. QUÉBEC, 2024). La figure 3.3 présente l'évolution de ces catégories par année pour les accidents mortels. On y observe la proportion de chaque genre d'accident parmi tous les accidents mortels annuellement.

FIGURE 3.3 – Proportion d'un genre d'accident parmi tous les accidents mortels de l'année en cours.

Les catégories « autre » , « collision avec animal » et « collision avec cycliste » présentent des valeurs nettement plus faibles que les autres catégories. On note malgré tout des pics en 2020 pour les collisions avec des objets fixes et les collisions impliquant des

cyclistes, ainsi que pour les accidents classés dans la catégorie « autre » ou les accidents sans collision en 2021. Un exemple de la catégorie « sans collision »serait le renversement d'un véhicule par lui-même.

Points d'inaptitude

Nous disposons également d'une variable indiquant le nombre de points d'inaptitude accumulés au cours des deux années précédant l'accident par les conducteurs impliqués. Nous espérions observer une augmentation du nombre moyen de points d'inaptitude accumulés par ces conducteurs pendant et après le déclenchement de la pandémie, mais comme le montre la figure 3.4, cette tendance n'est observable que pour les accidents graves et légers. Il est également surprenant de constater que les accidents mortels n'impliquent pas nécessairement des conducteurs ayant accumulé plus de points d'inaptitude que ceux impliqués dans des accidents graves.

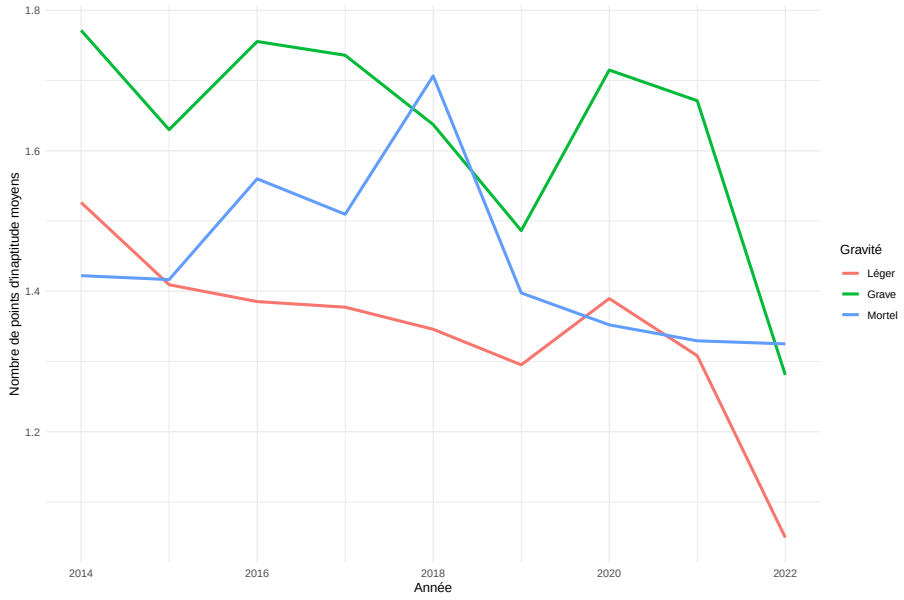

FIGURE 3.4 – Nombre moyen de points d'inaptitude d'un conducteur accumulés au cours des deux années précédant l'accident moyens par année et par gravité.

3.1.4 Dispositif de sécurité

Les données de la SAAQ fournissent également des informations sur l'utilisation des dispositifs de sécurité pour chaque usager de véhicules impliqués dans un accident. En créant un indicateur au niveau de l'accident pour identifier si au moins un usager n'utilisait pas correctement un dispositif de sécurité, nous obtenons les résultats présentés à la figure 3.5. Notons que les dispositifs de sécurité mal utilisés incluent les catégories suivantes : « inexistant », « non utilisé », « ceinture mal utilisée », « siège d'auto pour enfants mal utilisé », « casque mal ou non utilisé ». Dans tous les autres cas, même lorsque des données sur les dispositifs de sécurité n'ont pas été recensées, nous avons considéré que le dispositif de sécurité avait été porté correctement.

Cette hypothèse repose sur l'idée que, dans la majorité des cas, l'utilisation correcte des dispositifs de sécurité est la norme, et qu'une absence d'information peut souvent résulter d'un manque de collecte ou de signalement, plutôt que d'un usage incorrect. Cependant, si une proportion non négligeable des manquants correspond à des usages incorrects non recensés, l'ampleur réelle des problèmes liés aux dispositifs de sécurité pourrait être sous-estimée. Malgré cette limite, cette hypothèse permet d'intégrer tous les accidents dans l'analyse et d'éviter d'exclure les cas manquants, qui représentent une part importante des données.

Il semble y avoir des légers pics dans les pourcentages pour toutes les gravités d'accident en 2020 ou 2021. Cependant, il est à noter qu'il sera difficile d'intégrer cette covariable aux modèles en raison du manque de données. On ne recense que 7 539 accidents impliquant un dispositif de sécurité fautif sur environ 230 000 accidents survenus entre 2014 et 2022.

3.1.5 Cause principale de l'accident

Les policiers doivent spécifier les deux causes les plus probables de l'accident. Nous avons d'abord procédé à une reclassification des causes probables afin obtenir des catégories plus générales. Le Tableau 3.1 présente cette reclassification.

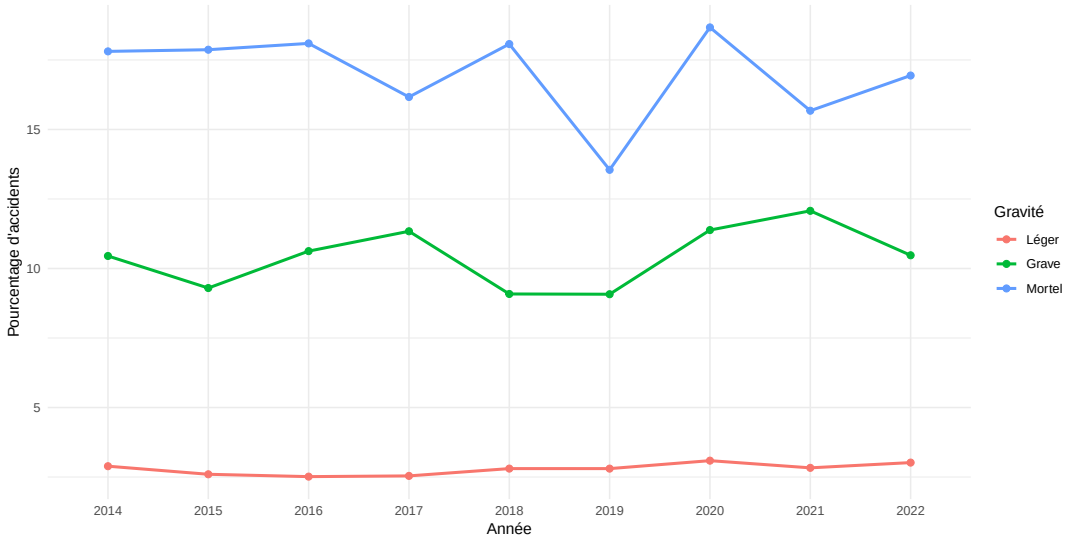

FIGURE 3.5 – Pourcentage d'accident impliquant au moins un usager n'utilisant pas bien son dispositif de sécurité en fonction de la gravité.

Catégories générales	**Causes**
Autres	Autres, Aucun facteur connu
Comportement risqué	Facultés affaiblies (alcool), Facultés affaiblies (médicament, drogues), Impatient, agressif, N'a pas fait un arrêt obligatoire, A passé sur un feu rouge, Autre comportement négligent, Excédait la vitesse permise, Conduite / vitesse imprudente, Arrêtait / tournait ou dépassait sans signaler, Suivait de trop près, Conduisait ou empiétait du mauvais côté de la voie, Circulait contrairement au sens unique, Reculait illégalement, Effectuait un dépassement interdit (ligne continue, courbe, zone de construction), Effectuait un dépassement dangereux, A dépassé ou croisé un autobus ou un minibus scolaire avec feux clignotants rouges, Faisait une course
Distraction	Inattention (n'a pas vu), Distraction due à l'équipement (radio, baladeur mp3, climatisation, chauffage, etc.), Distraction due à une autre personne à l'intérieur du véhicule (ex. : discussion, etc.), Autres distractions, Distraction due à un écran ou un terminal véhiculaire, Distraction due à un élément à l'extérieur du véhicule (autre accident, personne, objet), Utilisait un téléphone cellulaire
Défauts mécaniques	Freins défectueux, Crevaison, Direction défectueuse, Phares ou feux défectueux, Chargement non conforme, Attache de remorque défectueuse, Autres défauts mécaniques, Pneus non adaptés à la période hivernale ou non sécuritaire pour la conduite
Environnement routier	Visibilité obstruée, éblouissement, Mauvais état de la voie, Signalisation inadéquate, Éclairage insuffisant, Tracé inadéquat, Conditions météorologiques, Obstacles temporaires sur la route, Mauvais entretien hivernal, Animaux sur la route
Infraction au Code de la route	N'a pas cédé le passage, Virait à un endroit interdit, Était stationné incorrectement ou dans un endroit dangereux, A omis d'allumer ses phares ou d'en diminuer l'intensité, Changeait de voie
Rien à signaler	Rien à signaler
État du conducteur	Fatigue, sommeil, Malaise soudain

TABLEAU 3.1 – Reclassification des causes probables d'un accident en catégorie plus générale.

La figure 3.6 illustre la proportion de chaque cause d'accident parmi tous les accidents mortels de l'année en cours. Remarquons que les distractions atteignent un maximum en 2020.

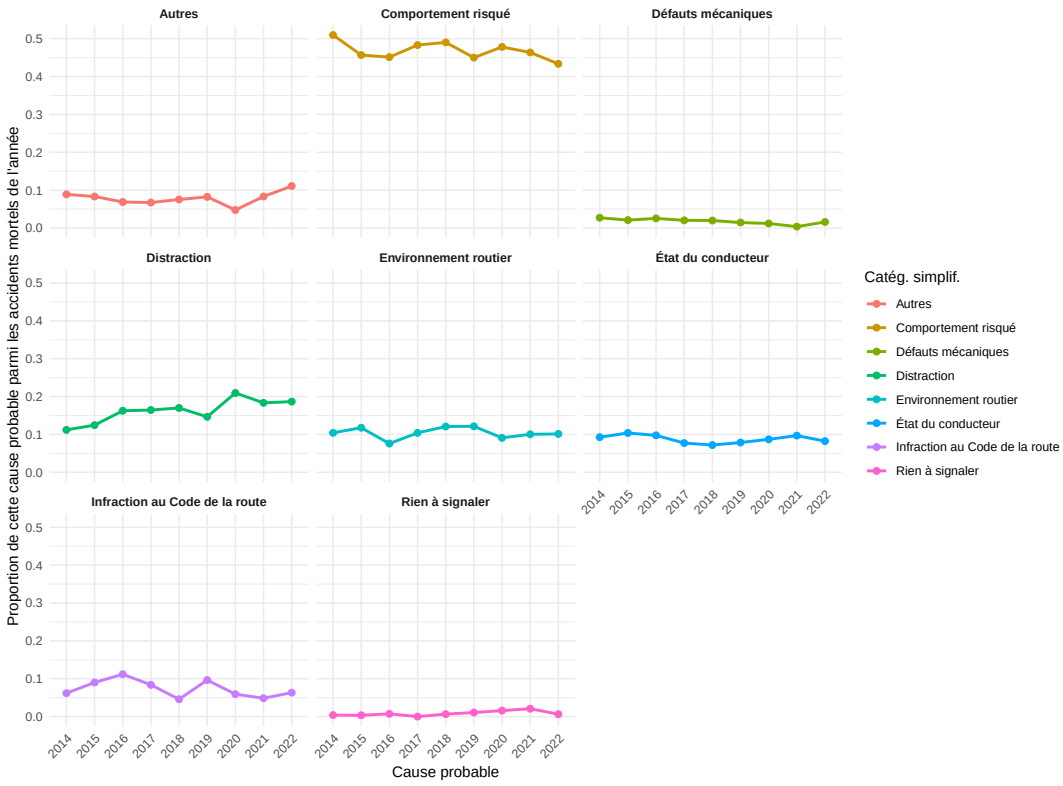

FIGURE 3.6 – Proportion de la cause probable de l'accident parmi tous les accidents mortels de l'année en cours.

3.1.6 Accidents impliquant des usagers vulnérables

Une de nos hypothèses pour expliquer pourquoi les accidents mortels n'ont pas diminué au même rythme que les autres types d'accidents avec la pandémie de COVID-19 est que le nombre d'accidents impliquant des usagers vulnérables a augmenté, et que ce type d'accident a beaucoup plus de chance d'être grave ou mortel. Les données de la SAAQ recensent trois types d'usagers vulnérables : les motocyclistes, les cyclistes et les piétons. La figure 3.7 montre l'évolution de la proportion d'accidents impliquant ces dif-

férents types d'usagers vulnérables par rapport à l'ensemble des accidents de leur gravité respective.

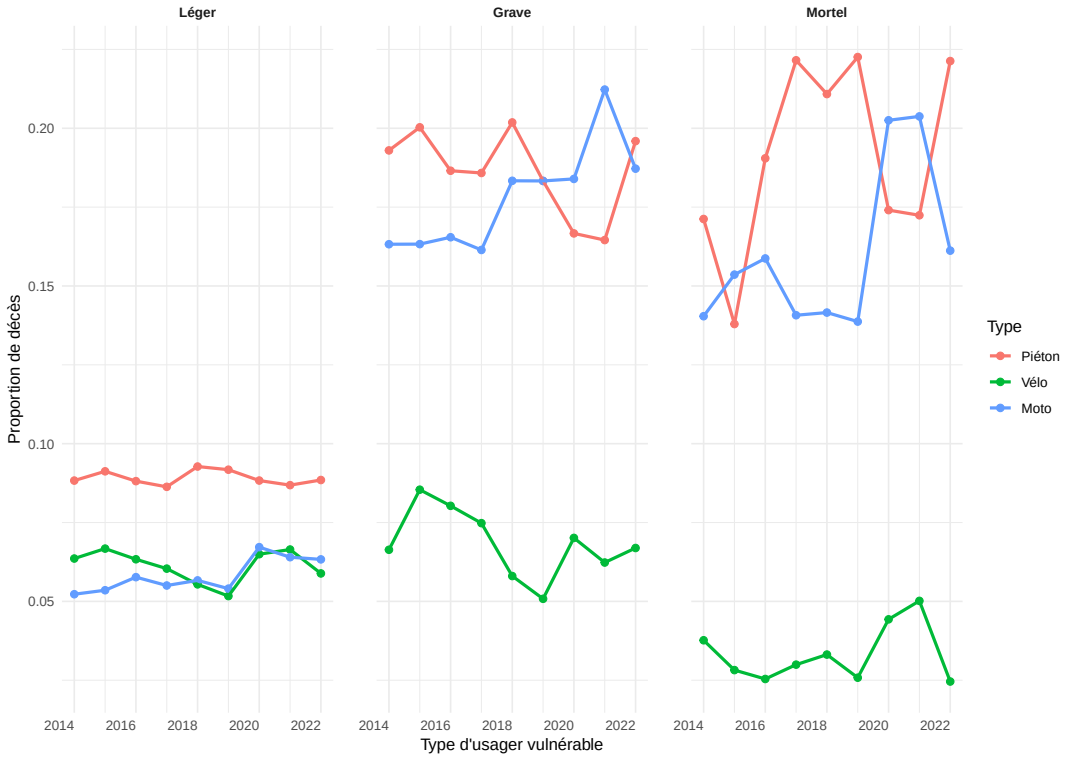

FIGURE 3.7 – Proportion d'accidents impliquant les différents types d'usagers vulnérables par rapport à l'ensemble des accidents mortels.

On observe, pour les accidents mortels, que la pandémie de COVID-19, qui frappe le Québec à partir de 2020, a entraîné une diminution de la proportion d'accidents impliquant des piétons en 2020 et 2021, mais une augmentation des accidents impliquant des motos. En revanche, en 2022, les tendances sont inversées. Pour cette raison, nous allons examiner la proportion de la somme des accidents mortels impliquant les trois types d'usagers vulnérables par rapport à l'ensemble des accidents mortels de l'année, comme présenté à la figure 3.8. On remarque aussi une certaine similarité des tendances des accidents graves et mortels impliquant piéton et moto.

Depuis le début de la pandémie, la proportion d'accidents impliquant au moins un des trois types d'usagers vulnérables est plus élevée qu'auparavant pour les accidents mortels et légers. Cela est dû, pour les accidents mortels, entre autres, à l'augmentation des ac-

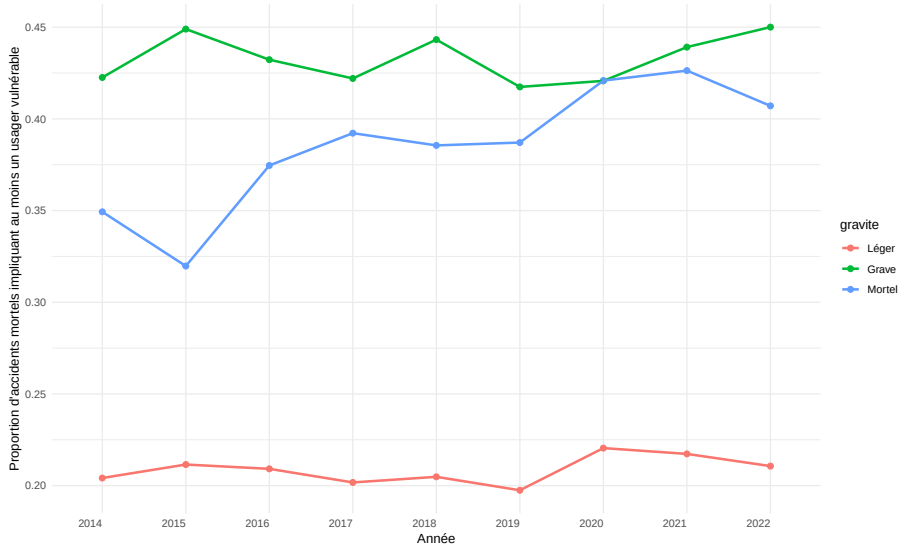

FIGURE 3.8 – Proportion d'accident impliquant au moins un usager vulnérable par rapport à l'ensemble des accidents de leur gravité, par type d'usager vulnérable et par année.

cidents impliquant des motos en 2020 et 2021, ainsi qu'à l'augmentation des accidents impliquant des piétons en 2022. On remarque aussi que la proportion d'accidents impliquant au moins un usager vulnérable augmente généralement depuis 2015.

3.1.7 Accidents n'impliquant qu'un seul véhicule

Une autre hypothèse est que les accidents n'impliquant qu'un seul véhicule ont pu augmenter avec l'arrivée de la pandémie. En effet, nous pensons qu'une augmentation des accidents n'impliquant qu'un véhicule pourrait être, entre autres, due à des conducteurs plus distraits. La figure 3.9 présente la proportion d'accidents n'impliquant qu'un seul véhicule par rapport à l'ensemble des accidents, en fonction de l'année et de la gravité.

La différence entre les accidents légers et les accidents graves/mortels est assez flagrante. Il y a en effet relativement plus d'accidents n'impliquant qu'un véhicule pour les accidents graves/mortels. Ensuite, on observe que la proportion d'accidents n'impliquant qu'un seul véhicule semble avoir augmenté depuis 2020, à l'exception d'une proportion maximale atteinte en 2019 pour les accidents mortels.

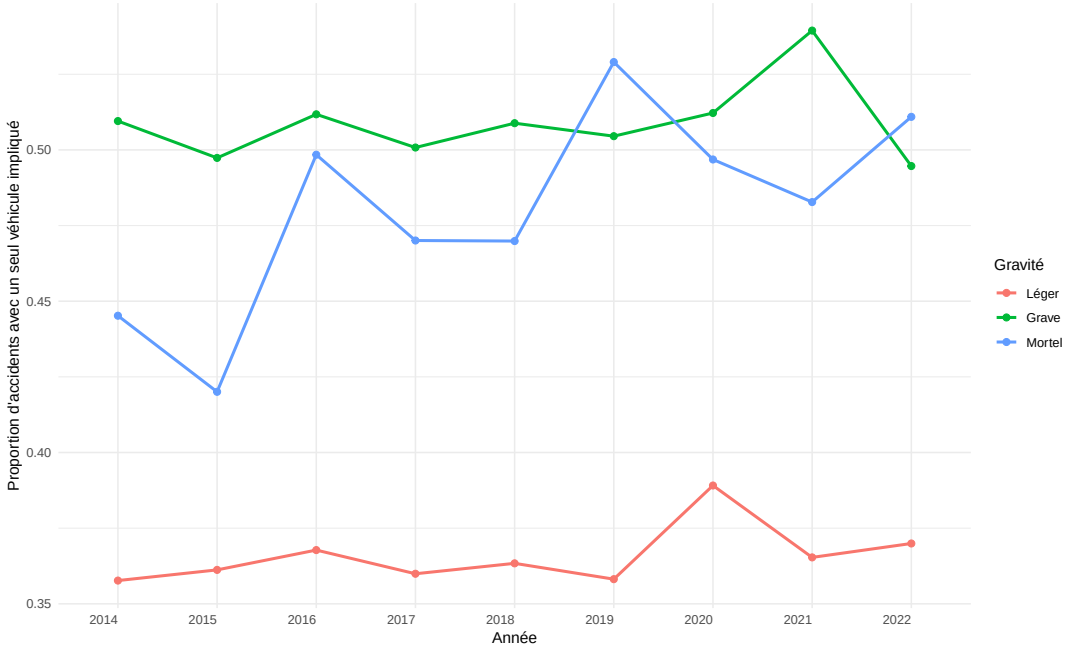

FIGURE 3.9 – Proportion d'accident n'impliquant qu'un seul véhicule par rapport à l'ensemble des accidents en fonction de l'année et de la gravité.

3.1.8 Données manquantes

Plusieurs autres variables sont disponibles. Cependant, bon nombre de ces variables présentent des valeurs manquantes, comme illustré à la figure 3.10. Rappelons que ce jeu de données est au niveau des conducteurs impliqués dans un accident. Nous avons conservé uniquement les accidents de gravité « léger », « grave » et « mortel » à partir de 2014. Ce jeu de données contient 402831 lignes.

Remarquons que la longitude et la latitude sont systématiquement absentes (manquantes 402831 fois), et que plusieurs autres indicateurs spatiaux sont également manquants. Ces absences peuvent être expliquées soit parce que les policiers n'ont pas rapportés ces informations, soit par les nouvelles lois de protection de données en vigueur au Québec qui empêchent la transmission des données. Certaines des variables manquantes, comme l'état de la chaussée, auraient pu être intéressantes à inclure dans l'analyse. Cependant, comme nous allons modéliser des décomptes par zones (municipalités et MRC), conserver davantage de variables aurait segmenté les données au point de se retrouver

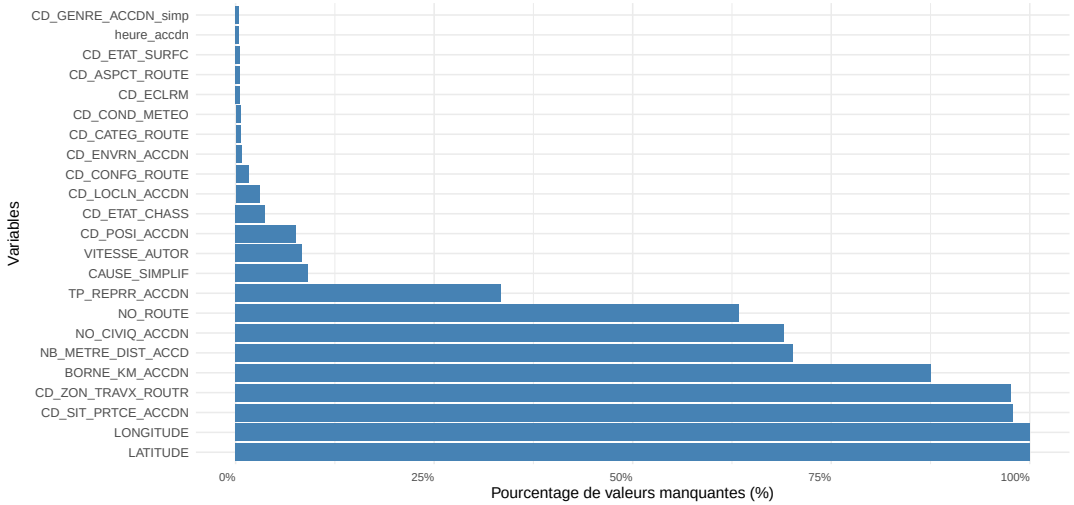

FIGURE 3.10 – Pourcentage de valeurs manquantes par variable.

avec trop de valeurs nulles par régions. Parmi toutes ces variables, il est tout de même intéressant de constater qu'il semble y avoir eu une augmentation des accidents mortels sur des routes sèches et lors de journées claires en 2020.

3.2 Données relatives aux accidents à Montréal

La Ville de Montréal met à disposition un sous-ensemble du jeu de données en ligne fourni par la SAAQ (`https://donnees.montreal.ca/dataset/collisions-routieres`). Ce sous-ensemble contient l'ensemble des accidents recensés sur l'île de Montréal, à l'exception de ceux survenus sur le réseau autoroutier. Grâce aux informations spatiales du jeu de données de la SAAQ, la Ville de Montréal calcule une géolocalisation de chaque accident. Il sera donc possible d'effectuer des analyses ponctuelles avec ces données.

3.3 Données externes

Les données relatives aux accidents nous fournissent beaucoup d'informations sur les accidents eux-mêmes, mais très peu sur d'autres facteurs démographiques, socio-économiques ou simplement spatiaux des environs de l'accident. C'est ici qu'entrent en

jeu d'autres sources de données pour approfondir nos analyses. Ces données externes proviennent du Gouvernement du Québec (P. d. QUÉBEC, s. d.), de *OpenStreetMap* (OSM) (OPENSTREETMAP, s. d.), de *cancensus* (CANCENSUS, s. d.), de la Communauté Métropolitaine de Montréal (CMM) (CMM, s. d.) et des débits de circulation estimés pour les routes et autoroutes gérés par le ministère des Transports et de la Mobilité durable (MTMD) (*Débit de circulation* s. d.).

Les données du Gouvernement du Québec utilisées sont celles des découpages administratifs des municipalités et des MRC. OSM est une carte du monde créée par des utilisateurs et libre d'utilisation sous une licence ouverte qui donne accès à une grande quantité d'informations spatiales à un niveau très fin. De plus, *cancensus* est une bibliothèque *R* qui permet d'accéder aux données de recensement quinquennal de Statistique Canada de 1996 à 2021. Dans ce travail, quand il sera question des données de *cancensus*, il sera en réalité question des données de recensement de Statistique Canada, mais extraites à l'aide de la bibliothèque *cancensus*. La CMM fournit des données très précises sur l'utilisation du sol dans la grande région de l'île de Montréal, ainsi que sur les routes jugées importantes.

Regardons maintenant plus en profondeur chacune de ces sources de données et les données spécifiques qui ont été extraites dans le cadre de ce projet. Les données spécifiques ont été choisies en s'inspirant des variables utilisées dans HUANG, ABDEL-ATY et DARWICHE, 2010 et SPYCHALA, 2023, qui utilisent des modélisations très semblables aux nôtres pour des accidents.

3.3.1 Découpages administratifs du Gouvernement du Québec

Le gouvernement du Québec met à disposition des *shapefiles* délimitant les municipalités et MRC de la province. Ces *shapefiles* seront éventuellement utilisés pour créer des voisinages nécessaires pour les modélisations zonales. Nous devrons ajuster quelque peu ces données pour les faire concorder avec celles de *cancensus*.

3.3.2 OpenStreetMap

Depuis sa fondation en 2004, OSM vise à créer une carte du monde libre et modifiable par tous. Ainsi, des données cartographiques détaillées et régulièrement mises à jour ont pu voir le jour. Aujourd'hui, plusieurs types de données y sont disponibles. Voici quelques exemples des données qu'on y trouve :

— Les routes, rues et autoroutes.

  — Panneaux d'arrêts, feux de circulation, etc.

— Les autres réseaux de transport (lignes de bus, de métro, pistes cyclables).

— Les zones d'usage du sol (parcs, zones résidentielles, industrielles, etc.).

— Les points d'intérêt (écoles, hôpitaux, restaurants, etc.).

Bien entendu, ces exemples ne sont qu'un aperçu de toutes les données d'OSM. Dans le cadre de ce projet de maîtrise, nous avons utilisé la bibliothèque *R osmdata* pour extraire les données suivantes.

— Intersections : Initialement, nous voulions extraire les intersections, mais il s'agissait d'un des rares repères spatiaux qui n'étaient pas référencés par OSM. Les feux de circulation et les panneaux d'arrêts ont donc été extraits comme variables substituts des intersections. Les feux de circulation et panneaux d'arrêt sont des données ponctuelles.

— Hôpitaux : Les hôpitaux ont été extraits dans le but de voir si la proximité d'un hôpital avait une incidence sur la gravité d'un accident. On cherche ici à voir si la proximité d'un hôpital résulte en des accidents moins graves puisque les blessés sont traités plus rapidement. Les hôpitaux sont des données ponctuelles.

— Routes : Le réseau de routes d'OSM est très complet et est constitué de différents niveaux allant des autoroutes jusqu'aux routes locales, donc des routes les plus importantes aux moins importantes sur le plan fonctionnel pour le trafic des véhicules à moteur. Le lien entre les routes d'OSM et les classifications routières québécoises

est obtenu à partir de `https://wiki.openstreetmap.org/wiki/Qu%C3%A9bec`
Les niveaux gardés dans ce projet sont :

— *motorway* : Autoroute, le plus haut niveau dans la hiérarchie du réseau routier. Il s'agit de route à deux vois où l'on peut conduire à plus de 100km/h.

— *trunk* : La prochaine catégorie des routes les plus importantes. Au Québec, il s'agit d'autoroutes faisant partie du réseau routier national (fédéral) mais qui n'est pas à chaussées séparées. Il n'y a donc qu'une ligne centrale jaune double entre les voies de sens opposés. Un bon exemple est l'autoroute 50.

— *primary* : La prochaine catégorie des routes les plus importantes dans le réseau routier d'un pays. Au Québec, il s'agit des routes dans l'intervalle 100-199, ou les routes principales dans des villes.

— *secondary* : La prochaine catégorie de routes les plus importantes dans le réseau routier d'un pays. Au Québec, il s'agit des routes dans l'intervalle 200-399, ou les routes importantes dans des villes.

— *tertiary* : La prochaine catégorie de routes les plus importantes dans le réseau routier d'un pays. Relie souvent des petites villes et villages.

Prétraitement pour la modélisation zonale

Les données sont brutes et devront être prétraitées. Les données ponctuelles, donc les feux de circulation et les panneaux d'arrêts, ainsi que les hôpitaux, ont simplement été comptées par municipalité et MRC. Pour ce qui est maintenant des routes, nous avons simplement sommé la longueur de routes de chaque type par municipalité et MRC. L'intersection entre les routes et les zones a été utilisée pour calculer cette longueur, de manière à ce qu'une route traversant plusieurs zones n'ajoute que la portion de route effectivement présente dans chaque zone. Cependant, nous n'avons pas pris en compte le nombre de voies par route. Pour chaque zone (municipalité ou MRC), il y a donc cinq mesures de longueur de route allant de *motorway* à *tertiary*. Nous avons aussi créé une variable *sommeRoute* où l'on somme la longueur de tous les types de routes par zone. Notons fi-

nalement qu'une autre source de données sur les routes est disponible : les données de débits de circulation (voir section 3.3.6). La longueur de route par zone a aussi été calculée à partir des données de débits journaliers moyens annuels (DJMA), et cette variable est nommée *routeQc*. *routeQc* correspond à peu près aux routes de type *motorway* jusqu'à *secondary*.

Les longueurs de routes et le nombre d'hôpitaux ont été normalisées en les divisant par la superficie (en km^2) de leur zone respective et en les multipliant par 10 pour être sur une échelle plus intuitive. Le nombre de feux de circulation et de panneaux d'arrêt a été combiné dans l'indicateur suivant nb. de feux de circulation $+ 0.5 \times$ nb. de panneaux d'arrêt, car il a été jugé que les panneaux d'arrêt sont placés à des intersections de moindre importance (*Tome V 2013 ; Signalisation routière - Volumes 1, 2 et 3* 2013). Cette nouvelle covariable, que nous nommerons *feuCircul_Stop*, a ensuite été divisée par la superficie, contrairement à la méthode utilisée dans HUANG, ABDEL-ATY et DARWICHE, 2010, qui divisait par la longueur totale des routes de la zone pour la normalisation. Nous avons préféré la division par la superficie afin de mieux gérer les valeurs extrêmes résultantes qui pourraient être obtenues en divisant par la longueur du réseau routier dans la zone.

Les tableaux 3.2 et 3.3 présentent des résumés des données brutes et prétraitées des municipalités et des MRC. Les premières rangées présentent les données brutes par zone, alors que les rangées suivantes présentent ces données avec leur ajustement respectif. Remarquons que ces dernières rangées sont très semblables pour les municipalités et les MRC.

Prétraitement pour la modélisation ponctuelle

La modélisation LGCP nécessite des valeurs de covariables sur l'ensemble de la région étudiée, et pas seulement aux endroits où des évènements/accidents ont eu lieu. Comme les données ponctuelles d'accidents sont disponibles sur l'ensemble de l'île de Montréal, nous devons avoir accès aux covariables partout sur l'île. Par conséquent, il a été nécessaire de compléter les données brutes d'OSM de manière à les rendre disponibles sur l'ensemble de l'île. Par exemple, une valeur indicative a dû être attribuée à des

Statistiques	Moyenne	Éc.-Type	Min.	Max.
Longueur des routes de type motorway (km)	3.4	11.8	0.0	173.3
Longueur des routes de type trunk (km)	1.7	8.5	0.0	148.0
Longueur des routes de type primary (km)	8.3	53.3	0.0	1814.5
Longueur des routes de type secondary (km)	9.6	38.5	0.0	989.9
Longueur des routes de type tertiary (km)	33.5	85.7	0.0	1570.5
Longueur des routes OSM sommées (km)	56.5	158.9	0.0	4090.2
Longueur des routes du MTMD (km)	24.1	50.0	0.0	1313.6
Nombre de feux de circulation	24.9	421.4	0	14643
Nombre de panneaux d'arrêt	46.3	345.7	0	10287
Nombre d'hôpitaux	0.2	1.7	0	56
Longueur motorway / Superficie · 10	0.5	1.5	0.0	14.3
Longueur trunk / Superficie · 10	0.1	0.6	0.0	9.2
Longueur primary / Superficie · 10	0.7	1.3	0.0	12.2
Longueur secondary / Superficie · 10	1.1	2.2	0.0	23.8
Longueur tertiary / Superficie · 10	2.7	3.3	0.0	36.3
Longueur routes OSM sommées / Superficie · 10	5.1	6.0	0.0	54.4
Longueur routes MTMD / Superficie · 10	2.6	3.5	0.0	32.5
Nb. de feux de circul. / Superficie	0.5	4.4	0.0	131.8
Nb. de panneaux d'arrêt / Superficie	1.7	9.0	0.0	125.0
(Nb. de feux de circul. + 1/2 · Nb. panneaux d'arrêt) / Superficie	1.4	7.6	0.0	158.1
Nombre d'hôpitaux / Superficie · 10	0.2	7.5	0.0	268.8

TABLEAU 3.2 – Tableau récapitulatif des variables extraites d'OSM et des routes du MTMD par municipalité. Les premières rangées présentent les données brutes par municipalité (longueur des routes, nombre de feux de circulation, panneaux d'arrêt, etc.), tandis que les rangées suivantes montrent ces données ajustées. Les superficies sont exprimées en kilomètres carrés. Les statistiques concernent les 1282 municipalités du Québec.

emplacements ne comportant pas initialement de feux de circulation.

Pour les covariables ponctuelles (feux de circulation, panneaux d'arrêt, hôpitaux), nous avons décidé d'utiliser le nombre d'observations de covariables se trouvant dans un rayon r autour de chaque point d'intérêt, par exemple le nombre de feux de circulation dans un rayon r autour de chaque point. La localisation des points d'intérêt est arbitraire, mais plus il y en a, mieux c'est, car cela permet d'augmenter la résolution spatiale de l'analyse. En ce qui concerne les covariables caractérisant des routes, on extrait plutôt la longueur de la route dans un rayon r autour de chaque point d'intérêt. Il est important de rappeler que, dans une modélisation LGCP, ces points d'intérêt correspondent précisément aux évènements/accidents et aux nœuds de la triangulation.

Étant donné que l'extraction de ces données à partir d'OSM est relativement coûteuse en termes de calcul, nous avons opté pour une extraction unique en suivant les étapes

Statistiques	Moyenne	Éc.-Type	Min.	Max.
Longueur des routes de type motorway (km)	44.2	50.8	0.0	258.4
Longueur des routes de type trunk (km)	21.8	46.2	0.0	221.0
Longueur des routes de type primary (km)	107.4	210.6	0.0	2027.4
Longueur des routes de type secondary (km)	126.0	148.8	0.0	1135.9
Longueur des routes de type tertiary (km)	437.7	322.0	4.2	1719.1
Longueur des routes OSM sommées (km)	737.0	576.4	72.8	4711.9
Longueur des routes du MTMD (km)	313.8	179.9	68.6	1492.5
Nombre de feux de circulation	324.1	1722.6	0	16749
Nombre de panneaux d'arrêt	601.6	1728.5	0	14558
Nombre d'hôpitaux	2.0	6.2	0	59
Longueur motorway / Superficie · 10	0.7	1.1	0.0	5.2
Longueur trunk / Superficie · 10	0.1	0.2	0.0	1.4
Longueur primary / Superficie · 10	0.6	0.7	0.0	4.4
Longueur secondary / Superficie · 10	1.0	1.7	0.0	14.7
Longueur tertiary / Superficie · 10	2.8	2.7	0.02	14.4
Longueur routes OSM sommées / Superficie · 10	5.2	5.7	0.1	38.1
Longueur routes MTMD / Superficie · 10	2.4	2.2	0.02	10.8
Nb. de feux de circul. / Superficie	0.8	3.7	0.0	33.6
Nb. de panneaux d'arrêt / Superficie	1.3	3.9	0.0	29.2
(Nb. de feux de circul. + 1/2 · Nb. panneaux d'arrêt) / Superficie	1.4	5.5	0.0	48.2
Nombre d'hôpitaux / Superficie · 10	0.03	0.1	0.0	1.2

TABLEAU 3.3 – Tableau récapitualtif des variables extraites d'OSM par MRC et des routes du MTMD. Les premières rangées présentent les données brutes par municipalité (longueur des routes, nombre de feux de circulation, panneaux d'arrêt, etc.), tandis que les rangées suivantes montrent ces données ajustées. Les superficies sont exprimées en kilomètres carrés. Les statistiques concernent les 98 MRC du Québec.

suivantes.

1. Les covariables ont été calculées autour des points d'intérêts illustrés à la figure 3.11. Les points rouges représentent les 263 accidents mortels survenus de 2012 à 2021. Les points en bleu sont 383 points répartis relativement uniformément sur le territoire, générés à l'aide d'une triangulation de Delaunay pour assurer une couverture homogène de la zone étudiée. Cette méthode permet de créer un maillage régulier tout en respectant la géométrie des frontières et des obstacles urbains. On se retrouve avec 646 points à partir desquels l'extraction sera effectuée.

2. Des fonctions *R* ont été créées pour calculer les valeurs des covariables aux 646 points utilisés. Le rayon r choisi est de 325 mètres. Il s'agit d'un choix arbitraire, mais basé sur ce qui a été fait dans SPYCHALA, 2023, où la modélisation LGCP utilise des cases de 650×650 mètres. Nous avons donc décidé de prendre un rayon

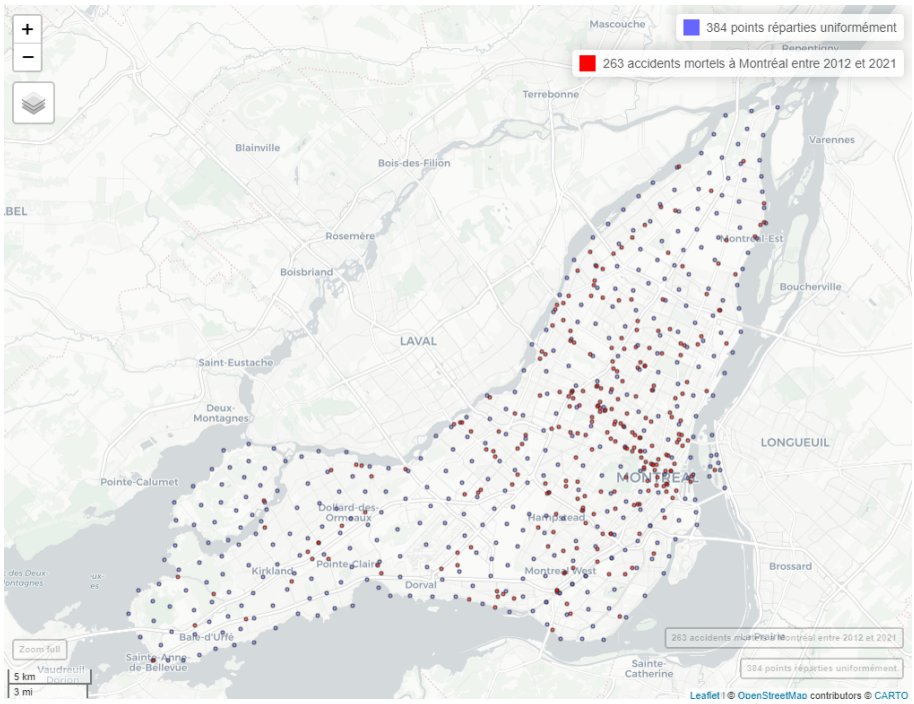

FIGURE 3.11 – 646 points sur l'île de Montréal d'où l'extraction des covariables sera faite.

correspondant à la moitié de 650 mètres.

3. Finalement, pour que les données soient disponibles peu importe où tombent les accidents et les nœuds de la triangulation, une fonction qui retourne simplement la valeur disponible la plus proche pour les valeurs manquantes a été créée. La figure 3.12 montre deux points avec des valeurs de covariables connues (en bleu) et un point dont la valeur doit être imputée (en rouge).

Le point en rouge représente une position où aucune valeur n'est disponible. Comme le point le plus proche avec une valeur connue est le point bleu situé en bas, le point rouge adoptera cette valeur. Ainsi, nous obtenons des données accessibles partout sur la carte. La figure 3.13 illustre la carte de la covariable des feux de circulation et des panneaux d'arrêts pondérés. Notons qu'ici aussi, nous avons combiné les nombres de feux de circulation et de panneaux d'arrêt et obtenu *feuCircul_Stop* de la manière suivante : nb. de feux de circulation $+ 0.5 \times$ nb. de panneaux d'arrêt.

Au tableau 3.4, on résume les covariables d'intersections et d'hôpitaux aux 646 points

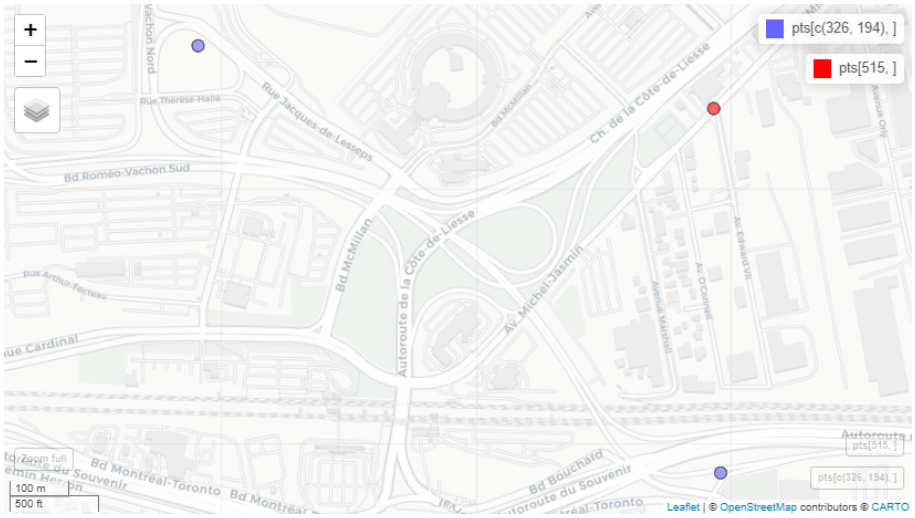

FIGURE 3.12 – Exemple d'imputation à la valeur la plus proche : le point en rouge, qui n'a pas de valeur initiale, prendra la valeur du point en bleu situé en bas, car c'est le point ayant la valeur la plus proche.

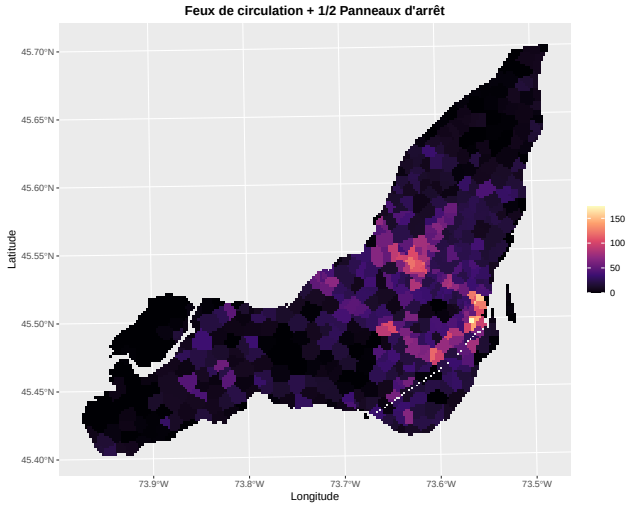

FIGURE 3.13 – Carte de Montréal illustrant les valeurs de la covariable combinant les feux de circulation et les panneaux d'arrêt, calculée selon les étapes décrites précédemment.

où l'extraction est effectuée.

Statistiques	Points	Moyenne	Éc.-Type	Min.	Max.
Nb. feu circul. + 0.5 · Nb. panneaux d'arrêt dans un rayon de 325m	646	22.4	25.5	0.0	154.0
Nb. hôpitaux · 10 dans un rayon de 325m	646	1.0	6.5	0	100

TABLEAU 3.4 – Tableau récapitulatif des variables extraites d'OSM sur l'île de Montréal, préparées pour la modélisation des processus ponctuels. Les statistiques sont calculées dans un rayon de 325 mètres autour des 646 points étudiés.

3.3.3 *cancensus*

Pour sa part, *cancensus* est un outil essentiel qui permet d'accéder aux données de recensement de Statistique Canada. Cette bibliothèque *R* facilite l'extraction des données des recensements 1996, 2001, 2006, 2011, 2016 et 2021, permettant ainsi d'analyser des informations démographiques et socio-économiques à différents niveaux géographiques, tels que les provinces, les MRC, les municipalités, et même par aire de diffusion, une petite unité géographique dont la population moyenne est de 400 à 700 habitants. Il y a 56 589 aires de diffusion au Canada. Dans les régions métropolitaines de recensement (RMR) dont le noyau compte au moins 50000 habitants, on recense aussi des secteurs de recensement (SR) ayant une population de moins de 7500 habitants (`https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/index-fra.cfm`). Les données de recensement sont cruciales pour compléter notre analyse des facteurs contextuels entourant les accidents. La bibliothèque *cancensus* propose près de 8000 variables de recensement. Parmi ces variables, nous avons sélectionné celles qui correspondaient à celles utilisées dans SPYCHALA, 2023, aux niveaux des MRC, des municipalités et des aires de diffusion :

— Population en 2021

— Densité de population par kilomètre carré

— Nombre de personnes de moins de 5 ans

— Nombre de personnes de 15 à 19 ans

— Nombre de personnes de 20 à 24 ans

— Nombre de personnes de 65 ans et plus

— Nombre de femmes

— Nombre de personnes ayant au moins un baccalauréat*

— Nombre de personnes n'ayant pas de diplômes postsecondaires*

— Nombre de chômeurs (15 ans et plus)*

— Revenu médian après impôt du ménage en 2020 ($)

— Nombre de bénéficiaires des prestations d'urgence et de redressement COVID-19 âgés de 15 ans et plus dans les ménages privés en 2020*

Veuillez noter que les « * » indiquent des variables pour lesquelles le recensement n'a été envoyé qu'à 25 % de la population, que Statistique Canada a ensuite interpolé pour obtenir une valeur à l'échelle du Canada.

Prétraitement pour la modélisation zonale

Les données de *cancensus* sont déjà disponibles au niveau des municipalités et des MRC. Il était donc permis de croire que le prétraitement serait plutôt facile. En effet, la variable « GeoUID » de *cancensus* contient déjà les codes géographiques des MRC et municipalités. Bien que ces codes concordaient en grande partie avec ceux utilisés dans les données d'accidents de la SAAQ, certaines zones différaient. Selon les données officielles de découpages administratifs du Québec, on recensait 1345 municipalités, alors que de *cancensus*, on identifiait 1282 municipalités. Pour les MRC, on avait le même problème avec 104 MRC des données officielles du Québec et 98 de celle de *cancensus*. Il est donc devenu primordial d'aligner le tout entre ces deux sources de données.

Alignement des régions Premièrement, notons que les deux municipalités disputées entre les provinces de Québec et Terre-Neuve-et-Labrador ont déjà été retirées du découpage administratif officiel.

Ensuite, il était évident que les données de municipalités officielles incluaient plusieurs régions situées sur l'eau alors que les municipalités de *cancensus* n'incluaient pas les municipalités sur des cours d'eau. De plus, ces régions avaient « aquatique » dans leur nom, En retirant ces municipalités « aquatiques » des données officielles, nous nous retrouvions avec 1295 municipalités, un nombre qui se rapproche des 1282 de *cancensus*.

Ensuite, les autres municipalités qui différaient encore ont été traitées individuellement. Les changements suivants ont été effectués :

1. La municipalité de la Morandière-Rochebaucourt n'est qu'une seule municipalité dans les données officielles du Québec, mais elle est séparée en deux municipalités (Morandière et Rochebaucourt) dans les données d'accidents de la SAAQ et de *cancensus*. Ces deux municipalités ont donc été fusionnées dans les données officielles.

2. Le code géographique de la municipalité de Notre-Dame-de-la-Salette dans *cancensus* a été modifié pour correspondre à celui des données officielles (de 82010 à 80087)

3. Les 14 municipalités restantes sont surtout des territoires autochtones du Nord-du-Québec, ou bien des petits territoires non organisés du sud du Québec. Nous avons simplement effacé ces municipalités des données officielles. Heureusement, aucun accident n'a été recensé dans ces municipalités.

Maintenant, pour ce qui est des MRC, notons d'abord qu'ici aussi, les deux MRC disputées entre les provinces de Québec et Terre-Neuve-et-Labrador ont déjà été retirées des données officielles. Les six MRC qui diffèrent entre les deux sources sont simplement des MRC qui sont unies dans *cancensus* mais séparées dans les données officielles, comme celle de Trois-Rivières présentée à la figure 3.14. Toutes ces MRC ont été réunies dans les données officielles, en prenant bien soin de mettre à jour le code de la MRC dans les données de la SAAQ également.

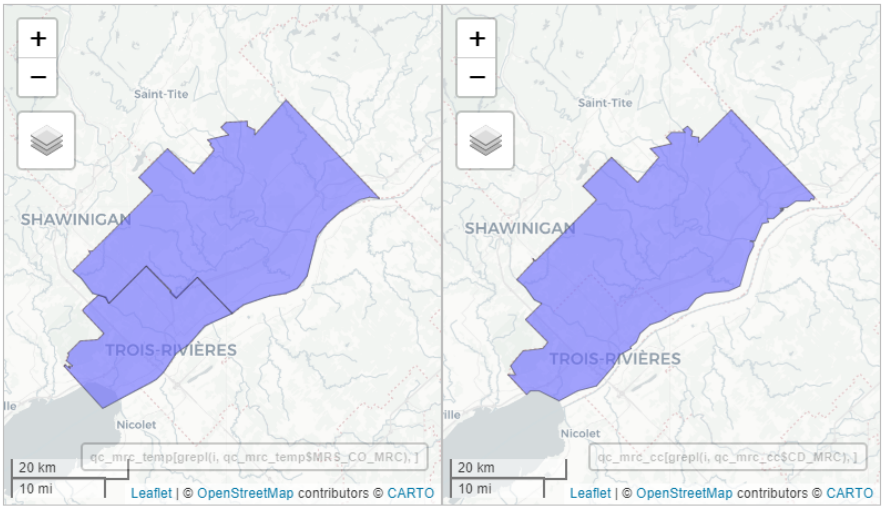

FIGURE 3.14 – Différence entre les découpages administratifs. À gauche : MRC provenant du découpages administratifs du Québec (P. d. QUÉBEC, s. d.) , où la MRC de Trois-Rivières est divisée en deux, soit la MRC de Trois-Rivières et celle de Les Chenaux. À droite : MRC provenant des données de *cancensus* (CANCENSUS, s. d.) , où la MRC de Trois-Rivières est déjà unifiée.

Prétraitement des données pour la modélisation zonale Tout d'abord, le nombre de personnes âgées de 15 à 19 ans et de 20 à 24 ans a été regroupé pour créer une covariable du nombre de personnes âgées de 15 à 24 ans. Le revenu médian par ménage a été divisé par 10000 et la densité de population (par km^2) a été divisée par 100. Toutes les covariables représentant le nombre de personnes d'une catégorie par zone ont été divisées par la population totale de la zone et multipliées par 100 pour obtenir le pourcentage de personnes appartenant à cette catégorie dans la zone. Il est important de noter que dans les rares cas où la population de la zone est égale à zéro (ce qui n'arrive que dans 89 municipalités), nous avons décidé d'utiliser la moyenne du pourcentage de cette catégorie dans les autres zones pour éviter que des pourcentages de zéros des covariables de population indiquent systématiquement une baisse du nombre d'accidents.

Les tableaux 3.5 et 3.6 présentent des résumés des données prétraitées des municipalités et des MRC. Remarquons que les indicateurs statistiques sont très semblables pour

les municipalités et MRC.

Statistiques	Moyenne	Éc.-Type	Min.	Max.
Centaines de personnes par km^2	1.3	4.7	0.0	64.8
% de moins de 5 ans	5.0	1.9	0.9	15.7
% de femmes	48.4	2.6	23.3	78.8
% de jeunes	8.7	2.5	1.8	25.5
% de 65 ans et plus	24.4	7.8	1.9	97.5
% de bacheliers et plus	10.4	7.0	0.7	62.5
% de personnes ayant un secondaire ou moins	19.4	5.9	3.3	46.0
% de chômeurs	3.8	2.1	0.6	26.7
Revenu médian/10000	5.3	2.5	0.0	12.7
% de bénéficiaires de prestations de COVID-19	21.7	3.2	10.5	54.1

TABLEAU 3.5 – Tableau récapitulatif des variables extraites de *cancensus* par municipalités. Les variables de population ont été converties en pourcentages, tandis que le revenu médian est exprimé en dizaine de milliers de dollars. Les statistiques concernent les 1282 municipalités du Québec.

Statistiques	N	Moyenne	Éc.-Type	Min.	Max.
Centaines de personnes par kilomètre carré	1.7	4.9	0.001	40.2	
% de moins de 5 ans	4.8	1.0	3.0	9.6	
% de femmes	49.9	0.9	48.2	52.1	
% de jeunes	9.2	1.4	6.7	15.1	
% de 65 ans et plus	24.0	5.2	9.1	33.9	
% de bacheliers et plus	11.6	5.0	5.3	30.4	
% de personnes ayant un secondaire ou moins	17.4	3.3	10.5	25.4	
% de chômeurs	3.4	0.9	2.0	7.1	
Revenu médian/10000	6.2	0.9	4.8	9.1	
% de bénéficiaires de prestations de COVID-19	21.6	2.0	17.6	26.6	

TABLEAU 3.6 – Tableau récapitulatif des variables extraites de *cancensus* par MRC. Les variables de population ont été converties en pourcentages, tandis que le revenu médian est exprimé en dizaine de milliers de dollars. Les statistiques concernent les 98 MRC du Québec.

Il est intéressant de constater que, même s'il y a un peu plus de 50% de femmes au Québec, la moyenne (non pondérée) des zones est inférieure à 50%.

Prétraitement des données pour la modélisation ponctuelle

Tel que mentionné brièvement plus tôt, les données de recensement extraites de *cancensus* sont, entre autres, disponibles au niveau des aires de diffusion (blocs de population de 400 à 700 habitants) et des secteurs de recensement (bloc de population de moins de 7 500 habitants). La figure 3.15 montre les deux niveaux pour l'île de Montréal.

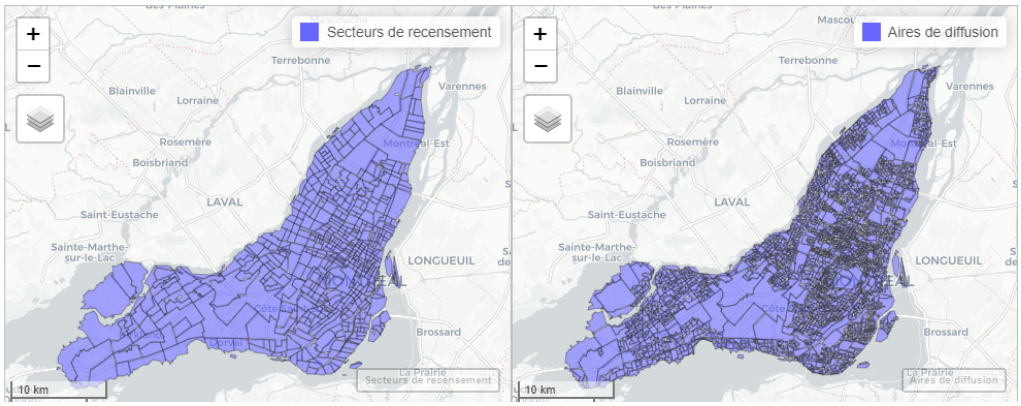

FIGURE 3.15 – À gauche, l'île de Montréal au niveau des secteurs de recensement. À droite, l'île de Montréal au niveau des aires de diffusion.

Dans ce projet, nous ne conserverons que le secteur de recensement (SR) comme niveau de covariable, étant donné que le niveau des aires de diffusion semblait trop fin, avec une variabilité trop élevée entre des points très rapprochés. Les covariables provenant de *cancensus* choisies sont présentées au tableau 3.7. Ces variables, sélectionnées en s'inspirant des travaux de SPYCHALA, 2023, visent à représenter l'environnement sociodémographique des différentes SR.

Statistiques	SR	Moyenne	Éc.-Type	Min.	Max.
Revenu médian par ménage / 1000 dans le SR	541	60.6	22.3	0.0	240.0
% de chômeurs dans le SR	541	5.5	1.8	0.0	15.0
% de jeunes (15 à 24 ans) dans le SR	541	11.7	4.2	0.0	41.8
% de vieux (65 ans et plus) dans le SR	541	16.4	7.5	0.0	58.5
Centaines d'habitants par kilomètre carré dans le SR	541	85.7	58.2	0.0	488.6
% de bénéficiaires de prestations de COVID-19 dans le SR	541	9.8	4.6	0.0	23.1

TABLEAU 3.7 – Tableau récapitulatif des variables extraites de recensements canadiens à l'aide de *cancensus* sur l'île de Montréal, préparées pour la modélisation ponctuelle. Les statistiques sont calculées en fonction des 541 secteurs de recensement (SR) à Montréal.

3.3.4 Communauté Métropolitaine de Montréal

La CMM a été créée en 2001 dans le but d'aider à planifier et coordonner les 82 municipalités qui composent la grande région de Montréal. Ces 82 municipalités comptent 4,2 millions d'habitants répartis sur plus de 4300 km^2. La CMM met à disposition plusieurs données, et celles qui seront utilisées dans ce projet sont :

— Utilisation du sol : L'occupation du sol de 2022 a été utilisée. Les différentes zones de la grande région de Montréal sont divisées en 11 grandes catégories. Seules les classes « Résidentielle », « Commerciale », « Bureau » et « Industrie » ont été retenues.

— Réseau artériel métropolitain : Le réseau artériel métropolitain est composé de routes considérées importantes pour le transport de gens et de marchandise, excluant les autoroutes. Il est lui-même composé de trois types de routes :

  — Voie de classe 1 : Routes assurant le déplacement entre les municipalités.

  — Voie de classe 2 : Routes servant de substitution aux autoroutes ou aux voies de classe 1.

  — Voie de classe 3 : Routes reliant le territoire aux autoroutes ou aux voies de classe 1.

Bien entendu, les données de la CMM ne seront utilisées que dans l'analyse ponctuelle que nous menons pour l'île de Montréal, puisque nous n'avons pas de données pour l'ensemble du Québec. Pour l'utilisation du sol, nous avons décidé d'utiliser comme covariable le pourcentage de la surface appartenant à la classe « Résidentielle » dans un rayon de 325 mètres autour des points d'intérêt. Nous avons également combiné les zones « Commerciale », « Bureau » et « Industrie » pour créer une deuxième variable d'utilisation du sol, en calculant à nouveau le pourcentage de surface appartenant à ces zones dans un rayon de 325 mètres. Nous avons conservé uniquement ces classes, car inclure toutes les classes aurait engendré un coût computationnel trop élevé. La distance de 325 mètres constitue un choix arbitraire, inspiré des travaux de SPYCHALA, 2023. Les routes du réseau artériel métropolitain, tous niveaux confondus, ont été combinées et extraites de la même manière que dans la section 3.3.2, donc en extrayant la longueur de routes dans un rayon de 325 mètres. Notons également que les mêmes étapes que celles présentées dans la section 3.3.2 ont été utilisées pour obtenir les 646 points où les données ont été extraites. Le tableau 3.8 montre un récapitulatif des données prétraitées.

Statistiques	N	Moyenne	Éc.-Type	Min.	Max.
% de zone résidentielle dans un rayon de 325 mètres	646	30.6	21.4	0.0	82.1
% de zones indus., comm., et bureau dans un rayon de 325 mètres	646	15.6	21.8	0.0	100.0
Kilomètres de routes dans un rayon de 325 mètres	646	1.0	1.0	0.0	5.0

TABLEAU 3.8 – Tableau récapitulatif des variables extraites des données provenant de la CMM et préparées pour la modélisation ponctuelle. Les statistiques sont calculées en fonction des 646 à Montréal.

3.3.5 Voisinage

Maintenant que les municipalités et MRC de *cancensus* et des données officielles du Gouvernement du Québec concordent, il ne reste qu'à créer des voisinages. Rappelons que nous avons besoin d'un voisinage pour la modélisation zonale. En effet, les modèles BYM modélisent des décomptes par zone en utilisant non seulement des covariables, mais également les décomptes des zones voisines. Cependant, en ajustant les données officielles de découpages administratifs, nous avons retiré toutes les municipalités aquatiques, ce qui a entraîné que le Québec se retrouve complètement séparé par le fleuve Saint-Laurent. Nous avons estimé qu'il était crucial de considérer comme voisines des zones reliées par des ponts. Nous avons donc superposé les routes et autoroutes dont la gestion incombe au MTMD au *shapefile* du découpage administratif. Les routes ont ensuite été élargies par l'ajout d'une zone tampon et ajoutées aux zones nécessaires pour obtenir le type de zones présenté à la figure 3.16.

En utilisant un voisinage contigu, nous obtenons le graphe de voisinage présenté à la figure 3.17.

Bien entendu, il y a beaucoup plus de liens dans le graphe des municipalités. Nous examinerons plus loin si ces liens apportent une meilleure précision aux modèles.

3.3.6 Débits de circulation estimés

Les débits de circulation estimés pour les routes et autoroutes dont la gestion incombe au MTMD sont disponibles. Plus de 4500 sites de collecte sont répartis sur l'ensemble du Québec et à l'aide de méthodes statistiques d'estimation, des débits sont estimés sur

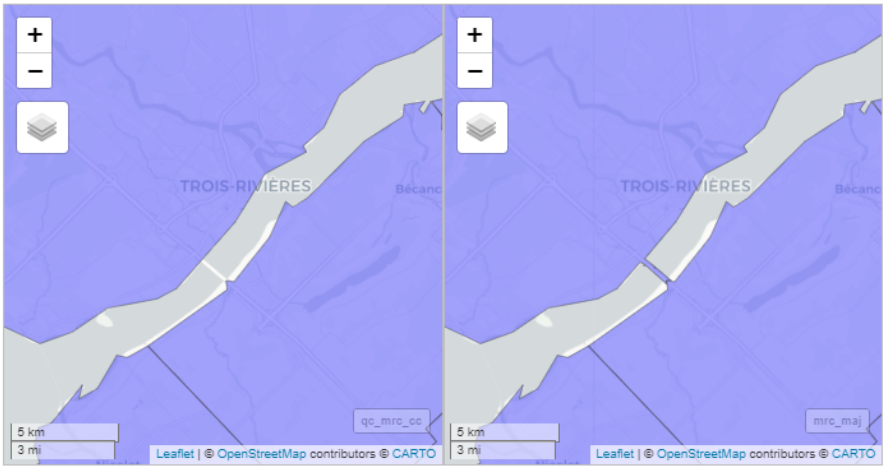

FIGURE 3.16 – À gauche, la MRC de Trois-Rivières et celle de Bécancour ne se touchent en aucun point, même si un pont les relie. À droite, ajustement de ces deux MRC pour qu'elles se touchent puisqu'elles sont reliées par un pont.

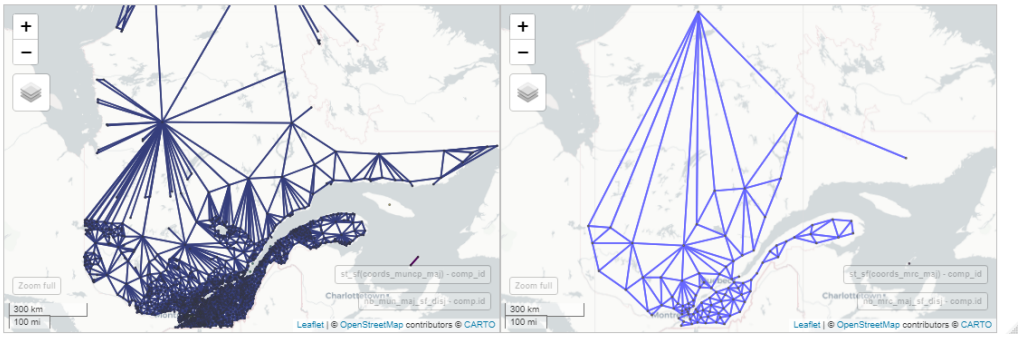

FIGURE 3.17 – À gauche, graphe de voisinage contigu des municipalités au Québec. À droite, graphe de voisinage contigu des MRC au Québec.

plusieurs autres routes. La variable qui est extraite pour ce projet est le débit journalier annuel moyen (DJMA). Le réseau de routes dont la gestion incombe au MTMD est présenté à la figure 3.18.

Remarquons que ces routes recouvrent surtout le sud de la province, soit là où une majorité de la population se trouve.

Exposition

L'exposition nécessaire pour les modèles BYM décrits à la section (2.5.3) est dérivée de la variable DJMA. Deux variables d'exposition ont été calculées. La première est naïve,

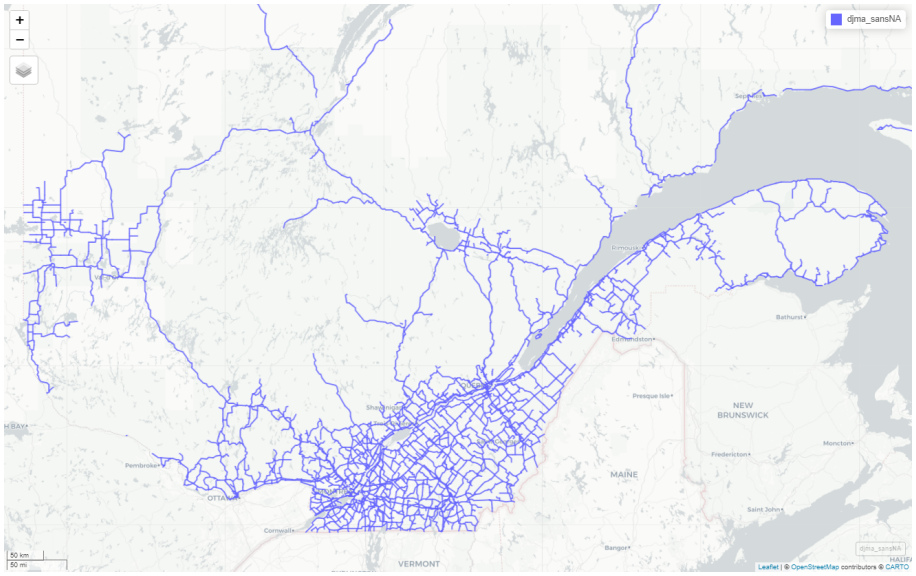

FIGURE 3.18 – Routes dont la gestion incombe au MTMD.

et la seconde, plus fidèle à la définition de quantité de déplacements décrite à la section 1.1.1.

Exposition naïve Cette approche simpliste s'appuie uniquement sur les données provenant des 4500 sites de collecte où les DJMA sont disponibles. Chaque route est ensuite associée à une région administrative en associant son centroïde à la région administrative correspondante. Pour chaque année, nous additionnons les DJMA de chaque région administrative. Nous choisissons de faire une somme plutôt que de calculer une moyenne des DJMA, car les régions administratives ayant plus de sites de collecte sont également celles avec un plus grand nombre de routes où des données sont collectées. Utiliser une moyenne entraînerait un effet de normalisation qui ne refléterait pas adéquatement l'importance et la quantité de ces routes. Cette exposition sera désignée comme l'exposition DJMA, par souci de concision.

Exposition plus fidèle à la quantité de déplacement Cette approche utilise les mêmes DJMA des 4500 sites de collecte, mais interpole ensuite aux autres routes où les données ne sont pas disponibles en se basant sur les débits des routes environnantes. La lon-

gueur de chaque section de route est ensuite utilisée pour calculer le nombre de véhicules-kilomètre. On se retrouve donc avec une mesure du nombre de véhicules-kilomètres parcourus (VKT), soit la longueur de la route multipliée par le nombre de véhicules ayant emprunté cette route. Encore une fois, nous avons sommé tous les débits par région administrative. Le travail pour interpoler tous ces débits a été décrit dans NETO, 2023. Cette exposition sera désignée comme l'exposition VKT, par souci de concision.

La figure 3.19 compare les deux types d'exposition, la méthode naïve et la méthode plus fidèle à la quantité de déplacement, que nous nommerons « VKT » pour plus de concision. Les deux expositions suivent des tendances très semblables, mais on remarque que l'exposition VKT est plus stable avant la pandémie. Enfin, il est important de noter que ces expositions sont collectées au niveau des régions administratives.

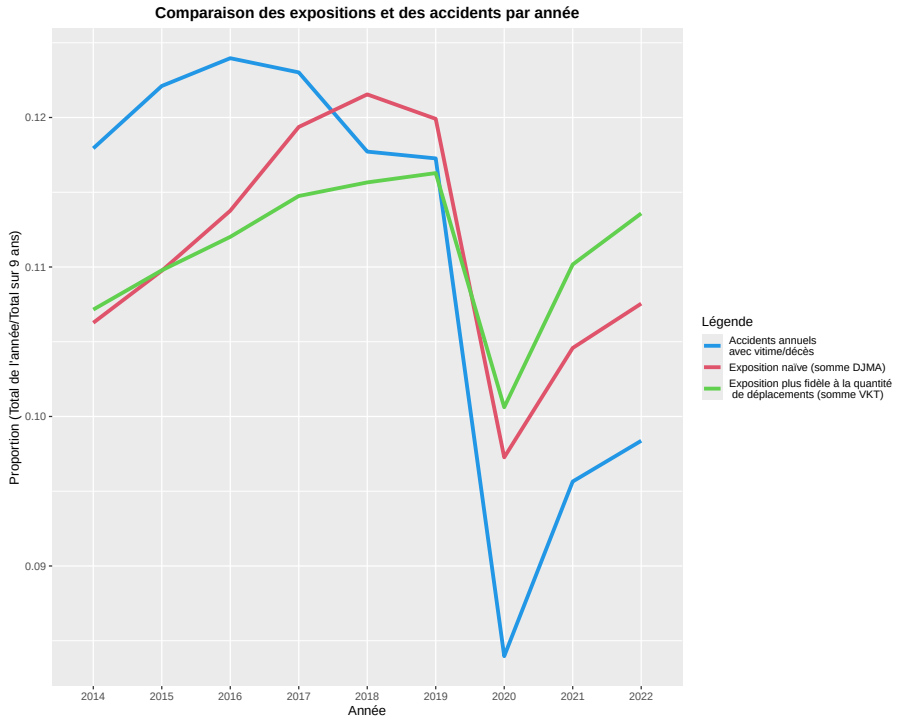

FIGURE 3.19 – Proportion des accidents survenus au cours d'une année donnée par rapport au total des accidents de 2014 à 2022. Les courbes vertes et rouges sont les proportions des expositions sommées sur tout le Québec pour l'année en cours par rapport au total des expositions de 2014 à 2022.

Ajustement aux MRC et municipalités Nous avons maintenant deux mesures distinctes de l'exposition, calculées par année au niveau des régions administratives. Cependant, nous souhaiterions affiner ces mesures à un niveau plus granulaire, c'est-à-dire au niveau des MRC ou des municipalités. On rappelle que dans les modèles BYM, l'exposition doit être accessible au niveau de l'agrégation spatiale. Pour ce faire, nous avons ajusté les données annuelles proportionnellement à la population obtenue grâce à *cancensus*. Par exemple, la municipalité de Montréal compte 1 762 949 habitants tandis que la région administrative en compte 2 004 265. L'exposition de la région administrative de Montréal s'élève à 12 212 000 passages. Par conséquent, l'exposition de la municipalité de Montréal en 2020 sera :

$$\begin{aligned}\text{exposition}_{\text{i=Montréal, j=2020}} &= \frac{\text{nb. habitants dans la municipalité de Montréal}}{\text{nb. habitants dans région la administrative de Montréal}} \\ &\quad \times \text{exposition}_{\text{Région administrative de Montréal en 2020}} \\ &= \frac{1762949}{2004265} \times 12212000.\end{aligned} \tag{3.1}$$

Notons qu'on a attribué une population de 1 aux municipalités sans population pour éviter des problèmes lors du calcul du logarithme de l'exposition.

Transformation des mesures d'exposition en nombre attendu d'accidents Une fois les mesures d'exposition disponibles pour chaque année et chaque zone, un dernier ajustement est nécessaire. Les modèles BYM nécessitent en effet des nombres attendus plutôt que simplement une exposition par zone pour que le risque relatif θ soit interprétable.

Pour transformer les mesures d'exposition en nombre attendu d'accidents, nous traitons individuellement les accidents de gravité légère, grave et mortelle puisque le nombre attendu pour chaque gravité d'accidents varie grandement. Nous calculons également un nombre attendu pour la somme des accidents graves et mortels.

Pour ces quatre types de gravité, il suffit de multiplier une mesure d'exposition (pour une zone particulière et une année spécifique) par le rapport entre la somme du nombre

d'accidents de cette gravité sur toutes les zones et années et la somme de l'exposition sur toutes les zones et années. Pour une zone i, une année j et une gravité g, on a :

$$E_{ijg} = \text{exposition}_{ij} \frac{\sum_{ij} \text{nbAcc}_{ijg}}{\sum_{ij} \text{exposition}_{ij}},$$

où E représente le nombre attendu d'accidents, nbAcc le nombre réel d'accidents et exposition ici est une mesure de la quantité de déplacements, donc dans notre cas, soit l'exposition DJMA ou VKT. Notons que cette exposition varie par zone i et année j, mais reste la même au-travers des gravités g. C'est de cette manière que sont calculé les nombre attendu d'accidents dans HUANG, ABDEL-ATY et DARWICHE, 2010 dans lequel on modélise modélise des décomptes d'accidents par zones à l'aide de modèle BYM.

Finalement, la figure 3.20 présente les nombres attendus d'accidents pour chaque gravité contre le nombre d'accidents réels. Notons que le nombre d'accidents attendus suit

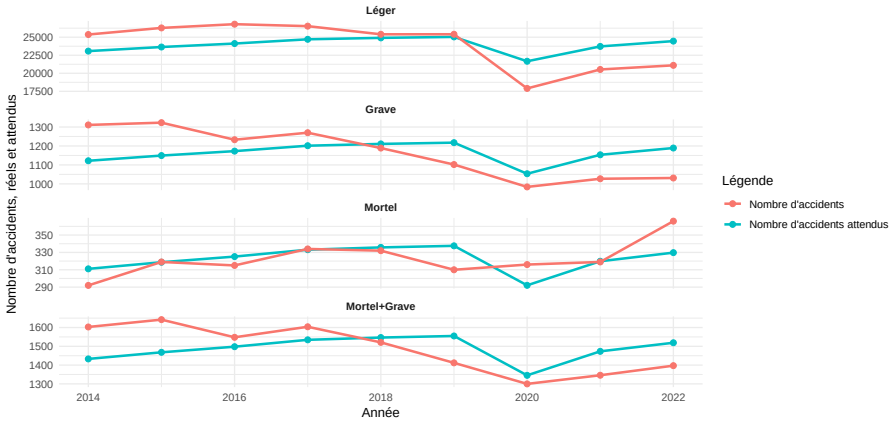

FIGURE 3.20 – Le nombre d'accidents, réels et attendus, pour les quatre types de gravités que nous utiliserons au courant de ce projet.

toujours la même tendance puisque le seul élément qui change pour chaque gravité est le facteur :

$$\text{nbAcc}_{ijg} = \sum_{i,j} \text{nombre d'accidents de gravité g}_{i,j}.$$

Remarquons aussi que la baisse du nombre d'accidents attendus en 2020 se trouvent entre celle des accidents mortels et graves/légers.

Nous allons maintenant analyser les données de la SAAQ et examiner si la pandémie de la COVID-19 a pu avoir eu un impact sur les accidents.

3.4 Jeux de données

Nous allons maintenant présenter les jeux de données finaux, en analysant la corrélation entre les variables et en présentant des résumés des variables retenues.

3.4.1 Jeux de données pour la modélisation zonale

Toutes les données ont déjà été prétraitées et sont prêtes à être utilisées, qu'il s'agisse de l'exposition, du voisinage ou des covariables. Il ne reste plus qu'à segmenter les données par période temporelle (année pour municipalités et MRC, et mois pour MRC) et par zone géographique. Ceci se fait en utilisant la bibliothèque *dplyr* de *R*. Cette bibliothèque permet de manipuler des jeux de données de manière efficace. Cependant, un problème persiste, surtout au niveau des municipalités, mais aussi avec les MRC : on a des zones pour lesquelles il n'y a pas d'accidents sur une période de temps donnée. Comme les modèles BYM sont des modélisations de décomptes, il faut bien entendu inclure les données là où le compte est de zéro. Cependant, en agrégeant simplement les données, on n'obtient pas d'informations pour les zones où aucun accident n'a été recensé. Il a donc été nécessaire d'ajouter des lignes correspondant à ces cas, sans quoi, on introduit de la troncature à gauche. Cette étape s'est révélée étonnamment complexe. Il a fallu séparer le jeu de données par gravité et année, puis compléter chacun de ces sous-jeux de données, ajouter des rangées pour les zones manquantes et remplir toutes les colonnes de ces nouvelles rangées. Cela a occasionné d'importants problèmes de mémoire avec *R*, et il a fallu optimiser le code pour minimiser l'utilisation de la mémoire.

Compléter chacun des sous-jeux de données s'est fait de la manière suivante. Les données provenant de OSM et de recensement canadiens sont déjà disponibles pour toutes les zones, puisque ces données ne sont pas temporelles, mais uniquement spatiales. Par conséquent, ces données sont simplement répétées d'année en année. Il aurait sûrement été plus judicieux d'utiliser des variables spatiales qui évoluent dans le temps, mais cela demandait trop de travail.

Notons que nous n'avons pas gardé les variables de points d'inaptitude et de vitesse

permise là où l'accident a eu lieu parmi les variables de la section 3.1.3. Les variables de type d'accident, de cause principale et de dispositif de sécurité n'ont pas été conservées dans ce jeu de données car elles auraient trop segmenté les données et nous avons déjà peu d'accidents graves et/ou mortels par zone. Finalement, nous avons aussi ajouté un indicateur pour la COVID-19. Pour les jeux de données segmentés par année, cet indicateur vaut 1 à partir de 2020, alors que pour le jeu de données segmenté par mois, cet indicateur vaut 1 dés mars 2020.

Corrélation

Bien entendu, avec la grande quantité de covariables disponibles, il est important de vérifier la corrélation pour éviter des problèmes de colinéarité dans les modèles. Les figures 3.21 et 3.22 montrent les cartes de chaleur des corrélations de Pearson pour les municipalités et MRC.

On remarque dans les deux *heatmaps* que les expositions (*nombreAccidents_VKT*, *nombreAccidents_DJMA*) et la variable cible (*nombreAccidents*) sont corrélées, ce qui était attendu. Avec le jeu de données qui segmente les accidents par municipalités et par année, on a peu de problèmes de corrélation entre les variables. En effet, il n'y a de forte corrélation (au-dessus de 0,8) qu'entre la longueur des routes de type *tertiary* et la somme des routes OSM, et entre la somme des routes OSM et la somme des routes gérées par le MTMD. De plus, il y a bien entendu une forte corrélation entre les variables du nombre de feux de circulation, de panneaux d'arrêts et de la somme pondérée des deux. Par conséquent, nous avons retiré la somme des routes OSM, le nombre de feux de circulation et de panneaux d'arrêt.

Par contre, il y a beaucoup plus de corrélation dans le jeu de données des accidents segmentés par MRC et par année ou mois. En effet, tous les types de routes OSM sont fortement corrélées soit entre eux, soit avec la somme de routes gérées par le MTMD. Nous ne garderons donc que la somme des longueurs des routes gérées par la MTMD. La densité de population est aussi fortement liée à la somme pondérée du nombre de feux de circulation et de panneaux d'arrêt et le nombre d'hôpitaux. Toutes ces variables

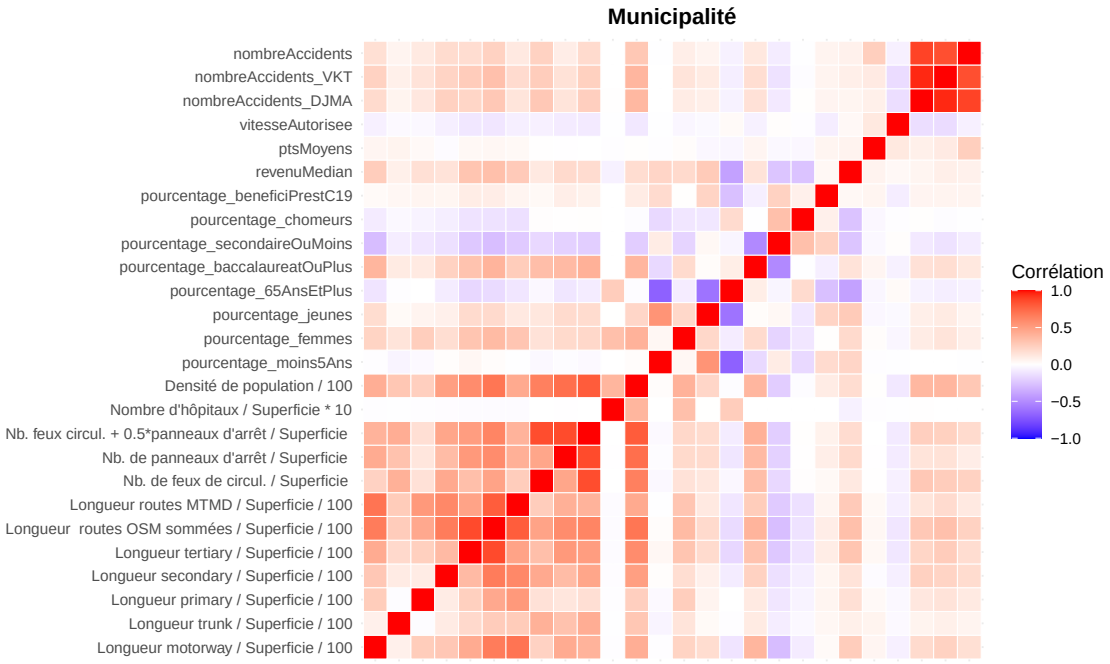

FIGURE 3.21 – Cartes de chaleur des corrélations de Pearson des jeux de données segmentant les accidents annuellement, par municipalités. Les trois premières variables sont les variables de réponse et d'exposition.

présentent également une forte corrélation aux expositions (*nombreAccidents_VKT*, *nombreAccidents_DJMA*) ce qui empêche leur inclusion dans les modèles. Nous avons donc retiré tous les types de routes OSM, leur somme, ainsi que le nombre de feux de circulation, de panneaux d'arrêt et leur somme pondérée, et finalement la densité de population. De plus, la variable *pourcentage_65AnsEtPlus* est fortement corrélée négativement aux variables *pourcentage_jeunes*, *pourcentage_moins5ans* et *revenuMedian*. Pour éviter des problèmes de multicolinéarité, cette variable a également été retirée.

Finalement, pour maintenir la cohérence des modèles, et ainsi pouvoir plus facilement les comparer entre eux, nous allons retirer les mêmes variables pour les municipalités et MRC. Ainsi, les variables conservées sont celles présentées dans le tableau 3.9.

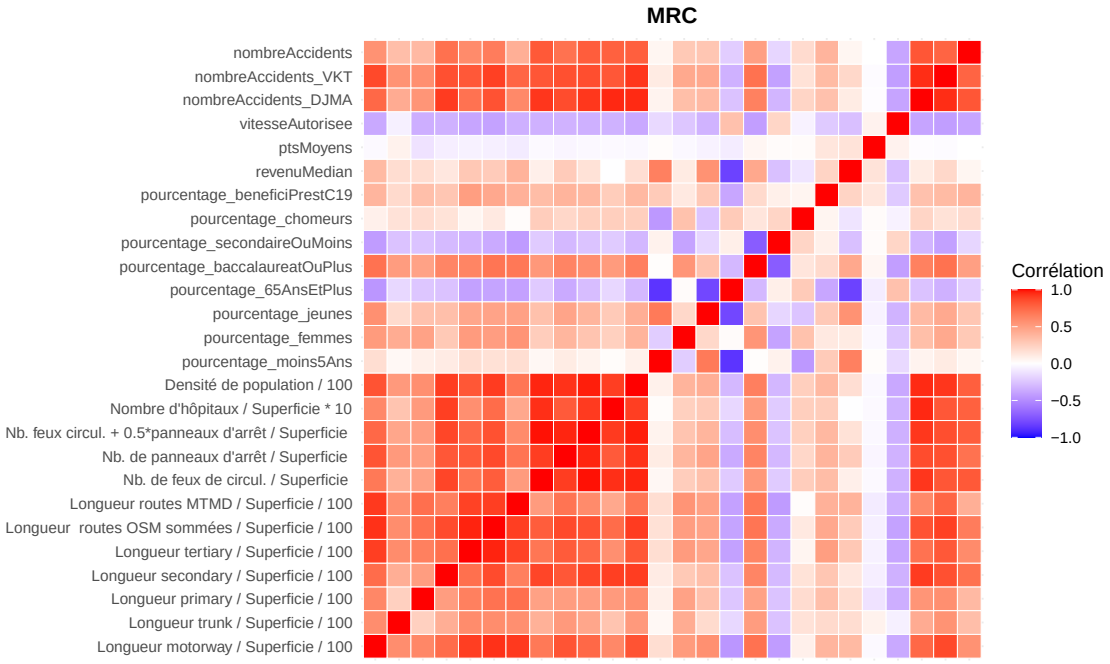

FIGURE 3.22 – Cartes de chaleur des corrélations de Pearson des jeux de données segmentant les accidents annuellement, par MRC. Les trois premières variables sont les variables de réponse et d'exposition.

Variable	**Description** (dans la zone)
densitéLinéaire_routesMTMD	Longueur de routes MTMD / Superficie / 100 dans la municipalité ou MRC
pourcentage_moins5Ans	Pourcentage de moins de 5 ans dans la municipalité ou MRC
pourcentage_femmes	Pourcentage de femmes dans la municipalité ou MRC
pourcentage_jeunes	Pourcentage de 15-24 ans dans la municipalité ou MRC
pourcentage_baccalaureatOuPlus	Pourcentage ayant un baccalauréat ou plus dans la municipalité ou MRC
pourcentage_secondaireOuMoins	Pourcentage de personnes ayant un secondaire ou moins dans la municipalité ou MRC
pourcentage_chomeurs	Pourcentage de chômeurs dans la municipalité ou MRC
revenuMedian	Revenu médian / 10000 dans la municipalité ou MRC
ptsMoyens	Points d'inaptitude moyens accumulés en 2 ans par les conducteurs impliqués dans un accident dans la municipalité ou MRC
vitesseAutorisee	Vitesse autorisée moyenne dans les zones où des accidents ont eu lieu dans la municipalité ou MRC
nombreAccidents_DJMA	Nombre d'accidents attendus avec la technique naïve (DJMA) dans la municipalité ou MRC
nombreAccidents_VKT	Nombre d'accidents attendus avec la technique plus fidèle à la quantité de déplacement (VKT) dans la municipalité ou MRC
nombreAccidents	Nombre d'accidents dans la municipalité ou MRC

TABLEAU 3.9 – Liste des variables conservées pour les modèles zonaux, accompagnées de leur description. On y retrouve aussi la variable dépendante *nombreAccidents*, et les variables d'expositions *nombreAccidents_VKT* et *nombreAccidents_DJMA*

Des statistiques descriptives de ces variables sont présentés aux tableaux 3.10 et 3.11. Ces tableaux récapitulatifs présentent des données déjà vues dans le chapitre 3, mais elles sont ici rassemblées en deux tableaux, un pour les municipalités (3.10) et l'autre pour les MRC (3.11) pour plus de clarté.

Statistique	Municipalités	Moyenne	Éc.-Type	Min	Max
densitéLinéaire_routesMTMD	1282	2.6	3.5	0.0	32.5
pourcentage_moins5Ans	1282	5.0	1.9	0.9	15.7
pourcentage_femmes	1282	48.4	2.6	23.3	78.8
pourcentage_jeunes	1282	8.7	2.5	1.8	25.5
pourcentage_65AnsEtPlus	1282	24.4	7.8	1.9	97.5
pourcentage_baccalaureatOuPlus	1282	10.4	7.0	0.7	62.5
pourcentage_secondaireOuMoins	1282	19.4	5.9	3.3	46.0
pourcentage_chomeurs	1282	3.8	2.1	0.6	26.7
revenuMedian	1282	5.3	2.5	0.0	12.7
ptsMoyens	1282	0.2	1.0	0.0	22.0
vitesseAutorisee	1282	78.4	6.6	0.0	100.0
nombreAccidents_DJMA	1282	0.3	2.3	0.000	76.4
nombreAccidents_VKT	1282	0.3	1.5	0.000	36.2
nombreAccidents	1282	0.3	1.0	0	36

TABLEAU 3.10 – Tableau résumé des variables de municipalités, segmentées par année et utilisées dans les modèles zonaux.

Statistique	MRC	Moyenne	Éc.-Type	Min	Max
densitéLinéaire_routesMTMD	98	2.4	2.2	0.02	10.8
pourcentage_moins5Ans	98	4.8	1.0	3.0	9.6
pourcentage_femmes	98	49.9	0.9	48.2	52.1
pourcentage_jeunes	98	9.2	1.4	6.7	15.1
pourcentage_65AnsEtPlus	98	24.0	5.2	9.1	33.9
pourcentage_baccalaureatOuPlus	98	11.6	5.0	5.3	30.4
pourcentage_secondaireOuMoins	98	17.4	3.2	10.5	25.4
pourcentage_chomeurs	98	3.4	0.9	2.0	7.1
revenuMedian	98	6.2	0.9	4.8	9.1
ptsMoyens	98	1.2	1.4	0.0	14.5
vitesseAutorisee	98	78.7	11.6	0.0	100.0
nombreAccidents_DJMA	98	3.3	8.9	0.1	86.9
nombreAccidents_VKT	98	3.3	5.5	0.3	41.1
nombreAccidents	98	3.3	3.7	0	37

TABLEAU 3.11 – Tableau résumé des variables de MRC, segmentées par année et utilisées dans les modèles zonaux.

Notons que les tableaux résumés 3.10 et 3.11 utilisent l'année 2020 pour le calcul des statistiques. Seules les variables *ptsMoyens*, *vitesseAutorisee* et *nombreAccidents* varient au fil des années. Toutes les autres variables demeurent constantes dans le temps.

3.4.2 Jeux de données pour la modélisation ponctuelle

Toutes les variables ont été prétraitées et sont disponibles pour l'ensemble de l'île de Montréal. Le tableau 3.12 présente les variables retenues pour la modélisation ponctuelle.

Variables	**Description**
feuCircul_Stop	Nombre de feu de circulation + 0.5 * Nombre de panneaux d'arrêt dans un rayon de 325m
hopitaux	Nombre d'hôpitaux dans un rayon de 325m (*10)
revenuMedian	Revenu médian par ménage / 1000 dans le SR
pourcentage_chomeurs	% de chômeurs dans le SR
pourcentage_jeunes	% de jeunes (15-24 ans) dans le SR
pourcentage_vieux	% de 65 ans et plus dans le SR
densitéPopulation	Centaines d'habitants par kilomètre carré dans le SR
pourcentage_prestaC19	% de bénéficiaires de prestations de COVID-19 dans le SR
routes	Kilomètres de routes dans un rayon de 325 mètres
pourcentage_zoneResidentielle	% de zone résidentielle dans un rayon de 325 mètres
pourcentage_zoneIndusComm	% de zones industrielle, commerciale, et bureau dans un rayon de 325 mètres

TABLEAU 3.12 – Variables retenues pour les modèles ponctuels, accompagnées de leur description.

Notons que les deux seules covariables présentant une forte corrélation (environ 0.75) entre elles sont *pourcentage_chomeurs* et *pourcentage_prestaC19*. Cette corrélation est naturelle, car il est intuitif que les chômeurs aient majoritairement bénéficié des prestations liées à la COVID-19.

Finalement, vous pouvez consulter les tableaux 3.7, 3.4 et 3.8 pour les statistiques descriptives des variables retenues.

Chapitre 4

Résultats

Cette section débute avec quelques analyses préliminaires et introduit les modèles de base, zonal et ponctuel, que nous utiliserons comme référence. Il sera ensuite temps de répondre aux trois grandes questions posées au début de ce projet.

4.1 Modélisations zonales de base

La modélisation zonale s'appuiera principalement sur des modèles BYM au niveau des municipalités et des MRC, avec pour objectif de modéliser l'effet de la COVID-19. Nous allons d'abord commencer par des analyses préliminaires suivies de modélisations plus approfondies.

4.1.1 Analyses préliminaires

Nous avons présenté plusieurs techniques d'analyses préliminaires dans les sections 2.3.4, 2.3.5 et 2.5.2. Ces analyses permettent de dégager une première idée, sans toutefois pouvoir en tirer des conclusions définitives. Commençons d'abord par regarder l'autocorrélation spatiale, et ensuite une estimation de l'intensité selon les différentes gravités.

Indice de Moran *I* global

Premièrement, nous examinerons l'indice *I* de Moran global pour le nombre d'accidents dans les MRC et municipalités selon différentes gravités. Étant donné que nous disposons de données annuelles sur le nombre d'accidents, nous avons additionné ces accidents dans chaque zone pour la période allant de 2014 à 2022, pour obtenir un seul compte dans chaque zone. De plus, cette approche permet d'avoir plus d'accidents par zone, améliorant ainsi la stabilité des comptes et renforçant la significativité statistique. Les résultats sont présentés dans le tableau 4.1.

Région	Gravité	Période	I de Moran global	Valeur-p
Municipalité	Léger	2014-2022	0.14	1.00e-32
MRC	Léger	2014-2022	0.21	2.92e-08
Municipalité	Grave	2014-2022	0.1	4.98e-19
MRC	Grave	2014-2022	0.15	3.84e-05
Municipalité	Mortel	2014-2022	0.09	1.51e-11
MRC	Mortel	2014-2022	0.1	0.027
				NA
Municipalité	Mortel	2014-2019	0.08	1.33e-08
Municipalité	Mortel	2020-2022	0.11	7.18e-15
MRC	Mortel	2014-2019	0.08	0.079
MRC	Mortel	2020-2022	0.11	0.016

TABLEAU 4.1 – Indice *I* de Moran global et les valeurs-p associés selon le type de régions, le type de gravité et différentes périodes temporelles.

Rappelons que l'hypothèse nulle stipule l'absence d'autocorrélation spatiale. Avec des valeurs-p inférieures à 5%, nous rejetons cette hypothèse et concluons qu'il existe une autocorrélation spatiale pour les accidents de 2014 à 2022 pour tous les types de gravités et de régions. En fait, il n'y a que pour les accidents mortels au niveau des MRC où l'on retrouve une valeur-p qui n'est pas quasiment nulle. Cela suggère que l'autocorrélation spatiale, bien que significative, est moins importante par MRC que par municipalités. C'était en quelque sorte attendu puisqu'une MRC regroupe plusieurs municipalités, ce qui peut lisser certaines variations spatiales locales. C'est probablement pour cela que l'autocorrélation spatiale est un peu plus faible (plus grande valeurs-p) au niveau des MRC.

On remarque aussi l'augmentation de la valeur-p pour les accidents mortels. Comme ce type d'accident est particulièrement pertinent dans notre étude, nous avons segmenté

les données en deux périodes : pré-COVID (avant 2020) et COVID (après 2020). Il est intéressant de noter que pour les MRC, le nombre d'accidents mortels avant la pandémie présente une valeur-p supérieure à 5%, ce qui signifie que l'autocorrélation spatiale n'est pas significative pour cette période. Il faudra donc vérifier que des modélisations spatiales ajoutent réellement quelque chose à l'analyse.

Cependant, les hausses observés des valeurs-p au niveau des accidents mortels pour les municipalités et MRC après 2020 pourraient simplement être attribuées à un plus petit échantillon (trois années d'accidents comparées à six avant 2020), réduisant ainsi la puissance statistique. Cela semble renforcé par des indices I très semblable pré et post-2020.

Diagramme de Moran

Les diagrammes de Moran, tant pour les municipalités que pour les MRC, sont présentés dans la figure 4.1. Les valeurs influentes sont déterminées à l'aide de la fonction *moran.plot* de la bibliothèque *spdep*. Le diagramme de Moran pour les municipalités montre que ces valeurs influentes (en rouge) sont réparties un peu partout au Québec, mais surtout autour des grands centres urbains. Cela était attendu, puisque nous avons utilisé les nombres bruts d'accidents, et non des valeurs normalisées par la superficie ou la population d'une zone. Pour les MRC, les valeurs influentes se concentrent principalement autour de Montréal et, dans une moindre mesure, de Québec. Cette observation est cohérente avec les valeurs-p légèrement plus élevées de l'indice de Moran global pour les MRC.

Indice de Moran I local

Le tableau 4.2 montre le nombre de régions (municipalités ou MRC) pour lesquelles l'indice I de Moran local du nombre d'accidents mortels de 2014 à 2022 est significatif, en d'autres termes, le nombre de régions montrant une autocorrélation spatiale significative. Ce tableau présente aussi un ajustement pour le taux de fausse découverte présenté dans

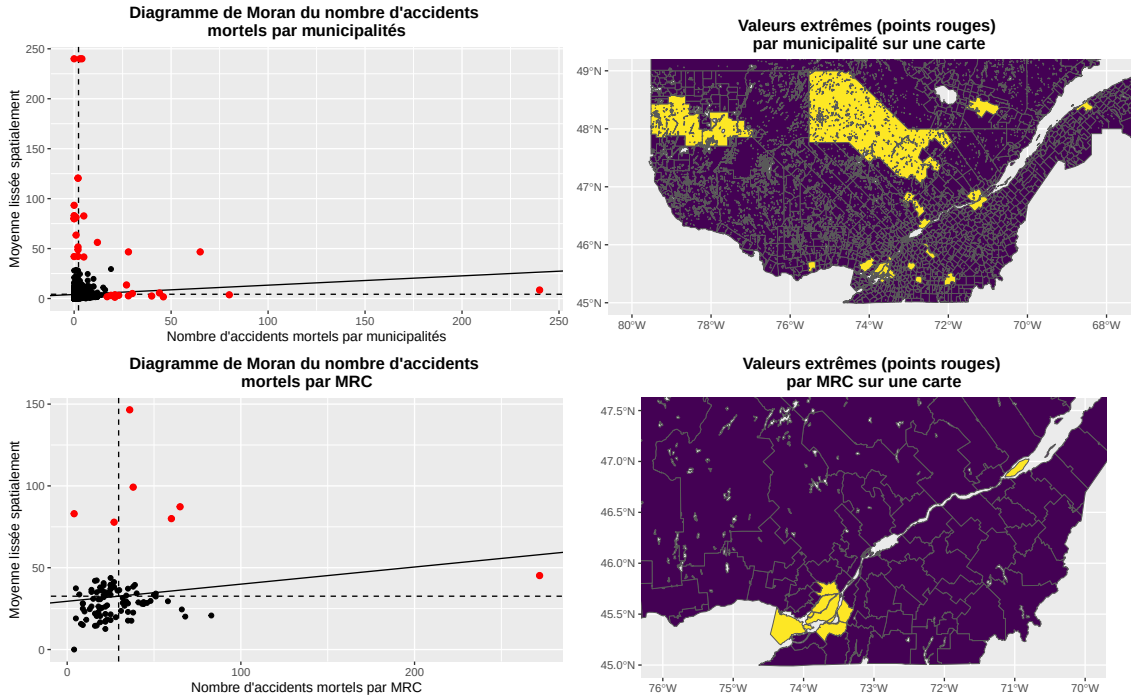

FIGURE 4.1 – Diagrammes de Moran (à gauche) du nombre d'accidents mortels de 2014 à 2022 dans les municipalités (en haut) et les MRC (en bas). Les valeurs influentes, indiquées par des points en rouge, sont en jaune sur les cartes à droite.

BENJAMINI et HOCHBERG, 1995.

Gravité	Région	*I* local significatif	FDR
Mortel	Municipalité	70	43
Grave	Municipalité	70	43
Léger	Municipalité	70	43
Mortel	MRC	6	5
Grave	MRC	7	6
Léger	MRC	6	6

TABLEAU 4.2 – Significativité du *I* de Moran local. La colonne « *I* local significatif » représente le nombre de zones pour lesquelles l'indice *I* de Moran local du nombre d'accidents mortels de 2014 à 2022 est significatif, c'est-à-dire le nombre de zones montrant une autocorrélation spatiale significative. La colonne « FDR » (pour *False Discovery Rate*) indique le nombre de zones pour lesquelles l'indice *I* reste significatif après ajustement pour le taux de fausses découvertes.

Logiquement, en ajustant pour le taux de fausse découverte, le nombre de régions présentant une autocorrélation significative diminue (ou reste le même). Pour rendre ces données plus visuelles, la figure 4.2 montre les régions présentant des autocorrélations

spatiales significatives pour le nombre d'accidents mortels. Ces régions sont classées en fonction de la nature de l'autocorrélation spatiale, positive ou négative. Pour ce faire, nous avons examiné la statistique de test Z : si elle est inférieure à -1.96, l'autocorrélation spatiale est considérée négative ; si elle est supérieure à 1.96, l'autocorrélation est considérée positive. Le seuil de 1.96 correspond au quantile de la loi normale permettant d'évaluer que l'autocorrélation est significative à 95%.

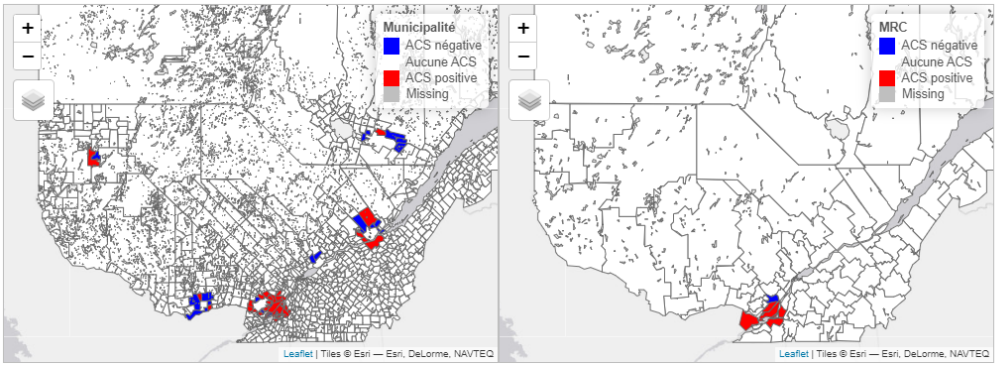

FIGURE 4.2 – Indices I de Moran locaux significatifs pour les municipalités et MRC au Québec. Les régions en rouge correspondent à une autocorrélation spatiale positive, tandis que celles en bleu indiquent une autocorrélation spatiale négative.

Pour les municipalités, l'autocorrélation spatiale est principalement observée autour des grands centres urbains tels que Montréal, Québec, Saguenay, Val-d'Or, Gatineau et Trois-Rivières. En revanche, pour les MRC, l'autocorrélation spatiale est concentrée autour de Montréal.

Enfin, nous avons démontré qu'il existe une autocorrélation spatiale pour les accidents légers, graves et mortels, bien que cette autocorrélation soit moins marquée pour les accidents mortels au niveau des MRC. Il est important de rappeler que ces analyses préliminaires n'incluent aucune covariable et que l'autocorrélation spatiale peut être induite par plusieurs éléments, comme des frontières administratives qui ne correspondent pas au processus spatial sous-jacent. Par exemple, des facteurs tels que l'urbanisation, les infrastructures routières, la densité de population ou encore des comportements de conduite particuliers peuvent également jouer un rôle dans la formation de cette autocorrélation spatiale. En particulier, dans notre cas, les régions plus peuplées sont celles présentant

de l'autocorrélation spatiale. Il sera essentiel d'intégrer des covariables pertinentes et de contrôler l'exposition dans les analyses futures.

4.1.2 Modèles zonaux de base

Nous allons maintenant ajuster plusieurs modèles BYM de base. Ces modèles serviront de référence pour prendre plusieurs décisions. Les décisions que nous aurons à prendre sont l'inclusion ou non de la région administrative du Nord-du-Québec, et l'utilisation d'une exposition naïve ou d'une exposition plus fidèle à la quantité réelle de déplacements. Étant donné un grand nombre de zéros présents dans les données des comptes d'accidents par zones, nous ajusterons en plus des modèles de Poisson, des modèles binomial négatif et *zero-inflated Poisson*. Comme plusieurs études mettent les accidents mortels et graves dans une seule catégorie, nous ajouterons aussi une catégorie des accidents mortels et graves combinés, en plus des catégories pour les types de gravités légers, grave et mortel.

Les modèles que nous ajustons incluent les covariables suivantes :

$$
\begin{aligned}
\textit{nombreAccidents} \sim\ & \\
& \textit{densitéLinéaire\_routesMTMD} + \\
& \textit{pourcentage\_moins5Ans} + \\
& \textit{pourcentage\_femmes} + \\
& \textit{pourcentage\_jeunes} + \\
& \textit{pourcentage\_baccalaureatOuPlus} + \\
& \textit{pourcentage\_secondaireOuMoins} + \\
& \textit{pourcentage\_chomeurs} + \\
& \textit{revenuMedian} + \\
& \textit{ptsMoyens} + \\
& \textit{vitesseAutorisee} .
\end{aligned}
\tag{4.1}
$$

Nous avons ajusté 96 modèles BYM différents pour prendre ces décisions. En effet, nous avons deux scénarios avec l'inclusion ou pas du Nord-du-Québec, deux scénarios pour l'exposition (VKT et DJMA), deux scénarios pour le type de zones (municipalité et MRC), quatre scénarios pour la gravité des accidents (légers, graves, mortels et mortels&graves combinés) et trois scénarios pour le type de distribution supposée pour la variable cible Y. La plupart des décisions seront prises en fonction du DIC.

Choix du type exposition et du type de distribution supposée pour la variable cible Y

Comme ces deux choix sont fait presque uniquement en se basant sur le DIC, le tableau 4.3 présente le DIC des 48 modèles incluant le Nord-du-Québec.

Modèle	Gravité	Région	Distribution	DIC_DJMA		DIC_VKT
BYM	Mortel	Munic.	Poisson	10684	>	10653
BYM	**Mortel**	**Munic.**	**BN**	**10635**	**>**	**10614**
BYM	Mortel	Munic.	ZIP	10688	>	10656
BYM	**Mortel**	**MRC**	**Poisson**	**3397**	**>**	**3370**
BYM	Mortel	MRC	BN	3401	>	3375
BYM	Mortel	MRC	ZIP	3400	>	3373
BYM	Grave	Munic.	Poisson	18677	>	18647
BYM	**Grave**	**Munic.**	**BN**	**18669**	**>**	**18636**
BYM	Grave	Munic.	ZIP	18680	>	18652
BYM	Grave	MRC	Poisson	4598	>	4575
BYM	**Grave**	**MRC**	**BN**	**4577**	**>**	**4545**
BYM	Grave	MRC	ZIP	4600	>	4578
BYM	Grave + Mortel	Munic.	Poisson	17011	>	16966
BYM	**Grave + Mortel**	**Munic.**	**BN**	**17006**	**>**	**16962**
BYM	Grave + Mortel	Munic.	ZIP	17015	>	16970
BYM	Grave + Mortel	MRC	Poisson	4839	>	4798
BYM	**Grave + Mortel**	**MRC**	**BN**	**4828**	**>**	**4780**
BYM	Grave + Mortel	MRC	ZIP	4842	>	4801
BYM	Léger	Munic.	Poisson	49739	>	49149
BYM	**Léger**	**Munic.**	**BN**	**48190**	**>**	**47956**
BYM	Léger	Munic.	ZIP	49742	>	49152
BYM	Léger	MRC	Poisson	9616	>	9241
BYM	**Léger**	**MRC**	**BN**	**8014**	**>**	**7921**
BYM	Léger	MRC	ZIP	9619	>	9244

TABLEAU 4.3 – Critère DIC en fonction de l'exposition. Les signes « > », « = » et « < » facilitent la comparaison entre les colonnes de DIC. Le tableau se divise en huit sous-groupes, chacun comprenant trois modèles. Dans chaque sous-groupe, les colonnes « Modèle », « Gravité », « Région » et « NQC » sont constantes, tandis que seule la colonne « Distribution » varie. Le meilleur DIC de chaque sous-groupe est indiqué en gras.

Il est évident que l'exposition notée VKT, plus fidèle à la quantité de déplacement,

produit systématiquement un meilleur DIC que l'exposition naïve, notée DJMA. On observe que la distribution optimale reste constante indépendamment de l'exposition, et que la distribution binomiale négative est presque toujours la meilleure. Pour ces raisons, nous continuerons avec l'exposition plus fidèle à la quantité de déplacement (VKT) et la distribution binomiale négative.

En ce qui concerne les accidents mortels dans les MRC, qui sont d'un intérêt particulier pour cette étude, nous avons comparé la distribution optimale des MRC, qui était celle de Poisson, avec la distribution binomiale négative. Afin de rester concis, nous ne présentons pas les coefficients des différents modèles, mais nous avons vérifié que les coefficients significatifs sont presque identiques entre les deux distributions. En effet, la majorité des coefficients sont pratiquement identiques, ce qui justifie la poursuite du projet en utilisant uniquement la distribution binomiale négative.

Enfin, pour éviter d'ajouter un tableau supplémentaire, notons simplement que les résultats demeurent exactement les mêmes lorsque nous excluons le Nord-du-Québec.

Inclusion ou non du Nord-du-Québec

Le DIC ne peut pas être utilisé directement pour décider de l'inclusion du Nord-du-Québec. En effet, en excluant cette région administrative, nous retirons également 57 municipalités et 1 MRC des jeux de données correspondants. Comme les jeux de données ne sont plus les mêmes, les valeurs du DIC deviennent non comparables. Selon la gravité et l'année, la région administrative du Nord-du-Québec représente entre 0,3 % et 1 % des accidents au Québec et 0,54 % de la population. Bien que la décision de l'exclure soit en quelque sorte arbitraire, nous avons choisi de retirer cette région en raison de sa vaste superficie — représentant 55 % de la superficie du Québec — et de sa faible population — seulement 1 % de la population totale.

Comparaison avec des modèles sans structure spatiale

On se retrouve maintenant avec huit modèles (quatre types de gravités et deux types de zones). Pour s'assurer que l'inclusion d'un effet spatial ajoute véritablement quelque chose à l'analyse, nous comparerons les DIC des modèles BYM à ceux d'un modèle très similaire, mais sans effet aléatoire spatial b_i. Les DIC sont présentés dans le tableau 4.4.

Gravité	Région	SS	DIC
Mortel	**Munic.**	**Avec_SS**	**10827**
Mortel	Munic.	Sans_SS	10845
Mortel	MRC	Avec_SS	3422
Mortel	**MRC**	**Sans_SS**	**3420**
Grave	**Munic.**	**Avec_SS**	**19018**
Grave	Munic.	Sans_SS	19047
Grave	MRC	Avec_SS	4598
Grave	**MRC**	**Sans_SS**	**4597**
Grave + Mortel	**Munic.**	**Avec_SS**	**17306**
Grave + Mortel	Munic.	Sans_SS	17357
Grave + Mortel	MRC	Avec_SS	4844
Grave + Mortel	**MRC**	**Sans_SS**	4843
Léger	**Munic.**	**Avec_SS**	**48760**
Léger	Munic.	Sans_SS	48777
Léger	MRC	Avec_SS	7993
Léger	MRC	Sans_SS	7993

TABLEAU 4.4 – DIC des modèles de base comparant l'inclusion ou non d'un effet spatial. La colonne « SS » (structure spatiale) indique si un modèle inclut (« AvecSS ») ou exclut (« SansSS ») une structure spatiale. Tous les modèles sont de type BYM, supposent une distribution négative binomiale pour les nombres d'accidents par zone, utilisent l'exposition VKT et incluent la région administrative du Nord-du-Québec.

Pour tous les modèles entraînés sur les municipalités, l'inclusion d'une structure spatiale améliore l'ajustement aux données. En revanche, pour les modèles ajustés au niveau des MRC, ce n'est pas le cas. En consultant la proportion de variance expliquée par la structure spatiale des modèles BYM au tableau 4.5, on confirme que la structure spatiale apporte effectivement peu de valeur aux modèles BYM entraînés sur les MRC.

En effet, la structure spatiale des modèles entraînés sur les municipalités explique entre 23% et 45% de la variance. Bien que ce pourcentage ne soit pas particulièrement élevé, il est demeuré non négligeable. Pour ce qui est des modèles entraînés sur les MRC, en excluant le modèle entraîné sur les accidents légers, moins de 2% de la variance est expliquée par la structure spatiale.

Gravité	Région	Prop. variance expliquée par SS
Mortel	Munic.	0.423
Grave	Munic.	0.236
Grave + Mortel	Munic.	0.272
Léger	Munic.	0.443
Mortel	MRC	0.011
Grave	MRC	0.018
Grave + Mortel	MRC	0.015
Léger	MRC	0.301

TABLEAU 4.5 – Proportion de la variance expliquée par la structure spatiale des modèles de base BYM pour les municipalités et MRC.

Nous concentrerons principalement notre analyse sur les modèles entraînés au niveau des municipalités. Cette approche est en ligne avec l'objectif du projet, qui ne cherche pas seulement à comprendre pourquoi les accidents mortels n'ont pas diminué, mais aussi à déterminer si ces accidents se produisent dans des lieux différents. De plus, cette orientation est cohérente avec les résultats des indices de Moran, qui indiquent une autocorrélation spatiale plus marquée au niveau des municipalités.

Coefficients des modèles de bases finaux

Au tableau 4.6, on présente les coefficients des modèles, au niveau des accidents mortels, graves et légers, pour les municipalités et les MRC (en incluant la structure spatiale). Nous ne conserverons pas les MRC dans toutes nos analyses, mais nous les montrons ici, puisqu'il s'agit du premier tableau de coefficients.

On remarque d'abord que les modèles entraînés en utilisant les municipalités comme voisinage n'ont pas les mêmes coefficients significatifs que ceux utilisant les MRC comme voisinage. Par exemple, *vitesseAutorisee* est significatif pour les accidents mortels dans le modèle basé sur les municipalités, mais ne l'est pas dans celui basé sur les MRC. Malgré cela, les coefficients semblent être semblables entre les modèles à différents niveaux d'agrégations. Interprétons maintenant certains des coefficients, en prenant comme référence le modèle des accidents mortels avec le voisinage municipal pour cette analyse.

1. *ptsMoyens* : Significatif pour les modèles basés sur les municipalités. Remarquons que plus l'accident est grave, plus ce coefficient est élevé. Le coefficient de 0.311

Coefficients	Modèle 1	Modèle 1.2	Modèle 1.3	Modèle 2.1	Modèle 2.2	Modèle 2.3
Région	Munic.	Munic.	Munic.	MRC	MRC	MRC
Gravité	Mortel	Grave	Léger	Mortel	Grave	Léger
(Intercept)	7.900 *	5.028 *	3.850 *	7.259 *	6.990 *	0.283
densitéLinéaire_routesMTMD	-0.031 *	-0.021 *	-0.007	-0.182 *	-0.113 *	-0.044
pourcentage_moins5Ans	0.046	0.041	0.025	0.034	-0.005	0.058
pourcentage_femmes	-0.138 *	-0.09 *	-0.051 *	-0.13 *	-0.141 *	-0.012
pourcentage_jeunes	-0.033	0.010	0.030 *	-0.039	-0.002	0.033
pourcentage_baccalaureatOuPlus	-0.001	0.015 *	0.009	0.015	0.023	0.020
pourcentage_secondaireOuMoins	-0.003	-0.008	-0.026 *	0.038	0.025	0.017
pourcentage_chomeurs	-0.015	-0.058 *	-0.084 *	0.074	0.081	0.037
revenuMedian	-0.226 *	-0.336 *	-0.279 *	-0.163 *	-0.123 *	-0.187 *
ptsMoyens	0.311 *	0.126 *	0.031 *	0.034 *	-0.012	0.098 *
vitesseAutorisee	0.013 *	0.024 *	0.01 *	0.002	0.006 *	0.004
DIC	10614	18636	47956	3375	4545	7921

TABLEAU 4.6 – Coefficients des modèles de base BYM entraînés sur les accidents mortels, graves et légers au niveau des municipalités à gauche, et MRC à droite. Les astérisques indiquent que le coefficient est significatif, c'est-à-dire que l'intervalle de crédibilité à 95% ne contient pas zéro.

signifie que pour chaque point d'inaptitude moyen supplémentaire accumulé par les conducteurs impliqués dans un accident dans la municipalité depuis 2 ans, le nombre d'accidents mortels augmente d'environ 36.5% $(\exp(0.311) \approx (1.365))$, en gardant toutes les autres variables constantes. Il n'est pas étonnant de voir cet effet. Cela suggère que les conducteurs avec plus de points d'inaptitude accumulés sont plus souvent impliqués dans des accidents mortels.

2. *pourcentage_femmes* : Il s'agit d'un coefficient qui se démarque par son amplitude parmi toutes les variables de pourcentage. Pour chaque augmentation de 1 point de % des femmes dans la municipalité, le nombre d'accidents mortels diminue d'un facteur d'environ 13% $(\exp(-0.138) \approx (0.871)\%)$, en gardant toutes les autres variables constantes.

Voici quelques autres faits intéressants. On observe qu'une densité linéaire des routes plus forte implique moins d'accidents dans tous les cas. De plus, on remarque aussi que le revenu médian plus élevé implique moins d'accidents, peut-être car ces quartiers plus riches ont de meilleures infrastructures avec moins de trafic de transit et de voies à grande vitesse.

4.2 Modélisations ponctuelles de base

Les seules données ponctuelles auxquelles nous avons accès sont des données d'accidents à Montréal de 2012 à 2021. Nous allons d'abord vérifier que les motifs ponctuels spatiaux (MPS) s'éloignent significativement d'un processus spatial complètement aléatoire (CSR) et ensuite estimer l'intensité de ces points dans différents cas.

4.2.1 Méthode de la fonction K de Ripley

D'abord, nous allons examiner si les MPS que nous avons s'éloignent d'un processus CSR, c'est-à-dire des motifs où les points sont indépendants les uns des autres. Si tel est le cas, il n'y aurait aucun intérêt à faire des analyses plus approfondies. La figure 4.3 montre les fonctions L (simple conversion de la fonction K rendant cette dernière linéaire) pour les MPS des accidents mortels, graves et légers ayant eu lieu sur l'île de Montréal de 2012 à 2021. Dans les trois cas, la fonction L observée diffère de la fonction L théorique. Cette dernière étant calculée sous l'hypothèse d'indépendance spatiale des points, correspondant à un processus CSR. Ces résultats montrent que, pour les trois gravités, les MPS ne suivent pas un processus CSR, ce qui suggère qu'il est pertinent de poursuivre l'analyse.

4.2.2 Estimation de l'intensité des accidents à Montréal

L'estimation de l'intensité des MPS des différentes gravités d'accidents à l'aide d'une fonction noyau est montrée à la figure 4.4. On y voit l'estimation des intensités des MPS sur toute la période de 2012 à 2021 à gauche, sur la période pré-COVID 19 dans la deuxième colonne et sur la période COVID-19 dans la troisième colonne. La quatrième et dernière colonne compare les MPS pré et pendant la COVID-19 pour les trois gravités. On y fait le ratio de l'intensité en ajustant les intensités pour qu'elle soit sur la même échelle (voir la section 2.4.2).

Tout d'abord, il est normal de constater des intensités plus faibles lorsque la période temporelle est plus courte. En examinant individuellement les accidents graves et légers,

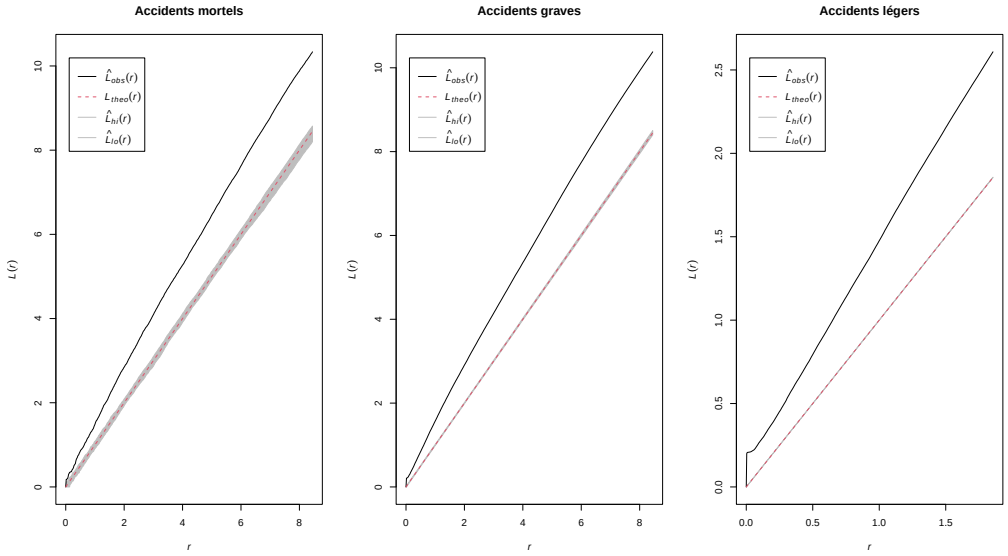

FIGURE 4.3 – Fonction $L(r)$ observée pour les accidents mortels, graves et légers à Montréal, comparée à la distribution théorique. Les courbes montrent $\hat{L}_{obs}(r)$, $L_{tho}(r)$, ainsi que les intervalles de confiance $\hat{L}_{lo}(r)$ et $\hat{L}_{hi}(r)$.

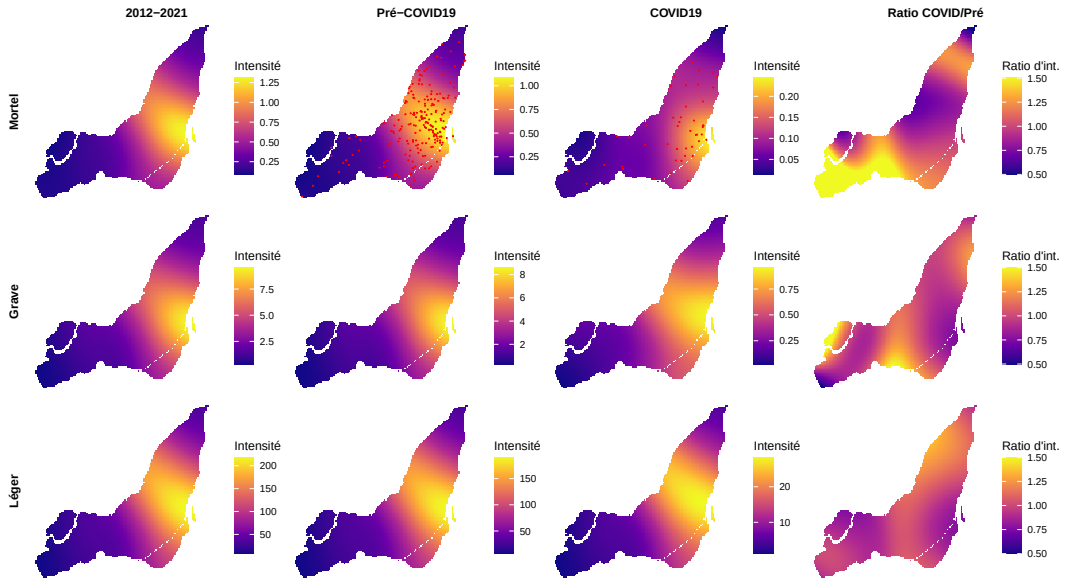

FIGURE 4.4 – Estimation des intensités à l'aide de fonctions noyaux pour les différentes gravités et périodes temporelles dans les trois premières colonnes. La dernière colonne présente un ratio ajusté des intensités pendant la COVID-19 par rapport aux intensités pré-COVID-19 pour les différentes gravités. Pour les accidents mortels, les points en rouge sont les localisations d'accidents.

les tendances pré-COVID et pendant la COVID-19 se ressemblent, ce qui semble être confirmé par le ratio qui reste proche de 1. Cependant, pour les accidents mortels, ceux-ci semblent moins concentrés dans un couloir passant au centre de l'île après la COVID-19 et davantage localisés au centre-ville. Cependant, en regardant le ratio ajusté pour les accidents mortels, on montre aussi une grande augmentation dans le sud-ouest de l'île.

La première hypothèse pour expliquer que ce couloir n'apparaît plus après la pandémie est que le nombre d'accidents mortels pendant la COVID-19 est simplement trop faible pour permettre une estimation fiable. Nous avons donc refait l'exercice pour les accidents mortels, mais cette fois avec des périodes égales. La période pré-COVID-19 utilisée s'étend du 13 mars 2018 au 31 décembre 2019, tandis que la période COVID-19 s'étend du 13 mars 2020 au 31 décembre 2021. Les graphiques sont présentés à la figure 4.5.

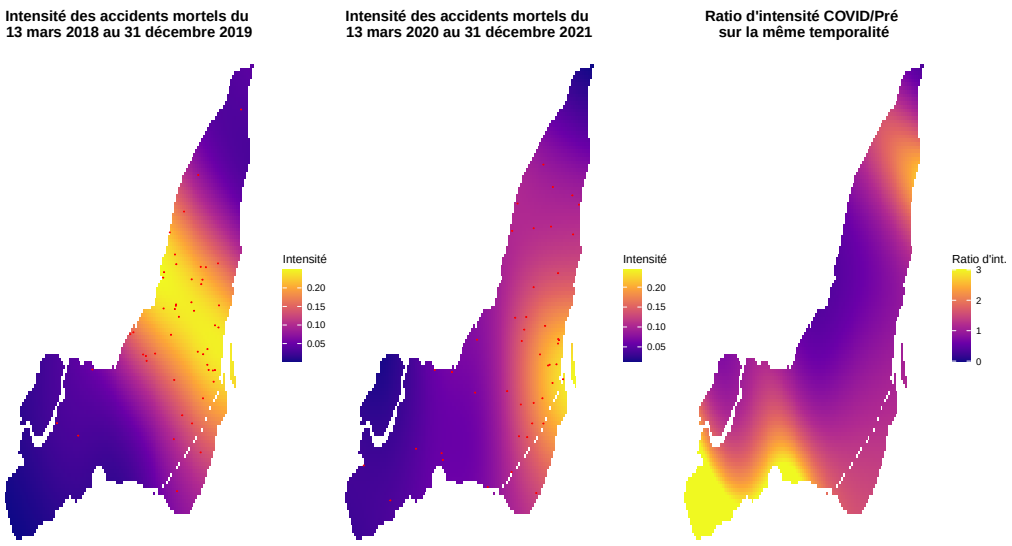

FIGURE 4.5 – Estimation des intensités des accidents mortels à l'aide de fonctions noyaux pour des périodes temporelles équivalentes. Les points en rouge représentent les emplacements des accidents. La dernière rangée montre un ratio ajusté des intensités depuis la début de la COVID-19 par rapport aux intensités pré-COVID-19.

Même pour des périodes temporelles équivalentes, les accidents mortels semblent moins concentrés au centre de l'île, se limitant principalement au centre-ville. En examinant de plus près le ratio d'intensité, on observe clairement une diminution de l'intensité

dans le nord de l'île (en violet foncé). Parallèlement, une augmentation est visible aux extrémités de l'île. Ces observations suggèrent une redistribution géographique des accidents mortels sur l'ensemble de l'île après la pandémie, accompagnée d'une concentration persistante au centre-ville. Dans les sections suivantes, nous tenterons de déterminer si ce changement de localisation est significatif et s'il peut être expliqué.

4.2.3 Modèles ponctuels de base

Les modèles ponctuels que nous avons examinés sont les modèles LGCP. Dans cette section, nous ajusterons des modèles LGCP contenant uniquement des covariables, donc sans processus spatial, puisqu'ils sont plus simples. Commençons par définir la triangulation que nous utiliserons pour tous les modèles ponctuels.

Tel que vu dans la section 2.4, nous avons besoin d'une triangulation de l'île de Montréal. On rappelle qu'il est préférable d'ajouter une extension autour de la triangulation pour minimiser les effets de bord. Cependant, comme on doit aussi avoir accès aux données/covariables à tous les noeuds de la triangulation, et que Montréal est une île, plusieurs de ces noeuds tombaient dans des cours d'eau. En pratique, nous avons donc décidé de limiter la triangulation aux limites de l'île (ou plutôt aux trois îles principales) de Montréal. La figure 4.6 montre la triangulation choisie pour la modélisation, même si cela génère des effets de bordure.

Les trois zones de couleur différente sont en réalité trois triangulations distinctes. Idéalement, la triangulation de l'île de Montréal (en vert) aurait inclus la zone délimitée par le canal de Lachine (en rouge), mais il était difficile de trouver des données géographiques précises sur les limites de Montréal. Celles qui ont été utilisées ici proviennent de *cancensus*. La triangulation a été définie de manière à ce que la longueur maximale d'une arête soit de de 2 km et que la distance minimale entre deux points soit de 1 km.

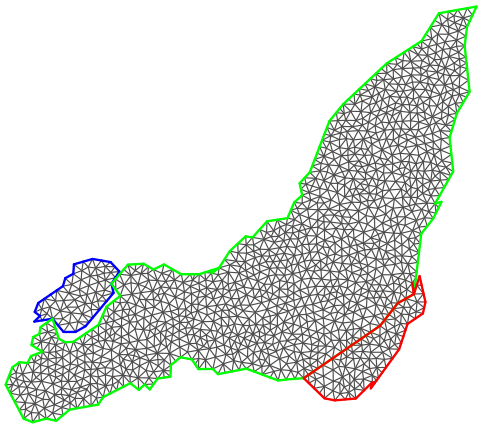

FIGURE 4.6 – Triangulation de l'île de Montréal. Les couleurs représentent les trois zones distinctes de la triangulation : l'île Bizard en bleu, l'île de Montréal en vert, et la zone délimitée par le canal de Lachine en rouge.

Modèles LGCP avec covariables uniquement

Tout d'abord, un modèle LGCP ne comprenant pas de processus gaussien n'est plus un LGCP puisque l'intensité modélisée ne sera plus stochastique. Il s'agit alors simplement d'un processus de Poisson. Cela implique également qu'on suppose une indépendance spatiale entre les accidents. En revanche, ces modèles sont plus simples et donnent une idée générale des effets des covariables. Comme ces modèles s'entraînent rapidement, ils seront utilisés pour déterminer s'il faut enlever la variable *pourcentage_prestaC19*, qui présente une forte corrélation avec la variable *pourcentage_chomeurs*.

Nous avons entraîné des modèles pour les accidents mortels et les accidents graves. Pour chacune de ces gravités, nous avons ajusté les modèles avec des données de différentes périodes temporelles. Ces périodes temporelles, accompagnées de leur nombre d'accidents, sont présentées au tableau 4.7.

En regardant les périodes temporelles égale, on constate que les accidents mortels sont passés de 49 à 45 accidents mortels, alors que les accidents graves sont passés de 277 à

Période	Nombre d'accidents mortels	Nombre d'accident graves
Complète : tous les accidents du 1e janvier 2012 au 31 décembre 2021.	263	1786
Pré-COVID : tous les accidents avant le 13 mars 2020.	218	1565
Période pré-COVID de même durée que la période COVID : tous les accidents du 13 mars 2018 au 31 décembre 2019.	49	277
COVID : tous les accidents aprés et incluant le 13 mars 2020	45	221

TABLEAU 4.7 – Nombre d'accidents graves et mortels sur l'île de Montréal selon différente gravité.

221 accidents. Cela correspond respectivement à des diminutions de 8% pour les accidents mortels et de 20% pour les accidents graves. Ainsi, la baisse des accidents mortels est moins prononcée que celle des accidents graves à Montréal également.

Maintenant que les périodes temporelles sont clairement définies, il ne reste qu'à ajuster les modèles. Les modèles que nous ajustons ont la formule suivante :

$$\begin{aligned} \textit{Intensité spatiale des accidents} \sim & \\ & \textit{feuCircul\_Stop} + \\ & \textit{hopitaux} + \\ & \textit{revenuMedian} + \\ & \textit{pourcentage\_chomeurs} + \\ & \textit{densitéPopulation} + \\ & \textit{routes} + \\ & \textit{pourcentage\_zoneResidentielle} + \\ & \textit{pourcentage\_zoneIndusComm} \\ & \textit{pourcentage\_jeunes} + \\ & \textit{pourcentage\_ainés} + \end{aligned} \tag{4.2}$$

Nous commencerons par ajuster les huit modèles (différentes périodes et gravités) avec toutes les covariables (voir tableau 3.12) et huit modèles avec toutes les covariables sauf *pourcentage_prestaC19*. Les DIC de ces modèles sont présentés au tableau 4.8.

Les modèles sans la variable *pourcentage_prestaC19* ont toujours un DIC inférieur à ceux l'incluant. Par conséquent, nous continuerons sans cette variable. Nous effectuerons

Gravité	Période	Avec ou sans *pourcentage_prestaC19*	DIC
Mortel	2012-21	Avec	-1952
Mortel	**2012-21**	**Sans**	**-1957**
Mortel	2012-19	Avec	-1611
Mortel	**2012-19**	**Sans**	**-1616**
Mortel	2018-19	Avec	-258
Mortel	**2018-19**	**Sans**	**-314**
Mortel	2020-21	Avec	-302
Mortel	**2020-21**	**Sans**	**-306**
Grave	2012-21	Avec	-13482
Grave	**2012-21**	**Sans**	**-13483**
Grave	2012-19	Avec	-11755
Grave	**2012-19**	**Sans**	**-11756**
Grave	2018-19	Avec	-2079
Grave	**2018-19**	**Sans**	**-2080**
Grave	2020-21	Avec	-1693
Grave	**2020-21**	**Sans**	**-1695**

TABLEAU 4.8 – DIC des modèles ponctuels avec covariables uniquement. On y compare les modèles avec et sans la variable *pourcentage_prestaC19* pour les accidents graves et mortels.

l'analyse plus approfondie de ces modèles dans la question de recherche sur la localisation des accidents (section 4.3.3).

4.3 Questions de recherche

Répondons maintenant aux questions de recherche. Ces questions seront :

1. Existe-t-il un effet COVID-19 significatif sur les accidents de la route au Québec ?
2. Quels facteurs expliquent l'augmentation significative des accidents mortels après 2020 ?
3. La localisation des accidents sur l'île de Montréal a-t-elle changée avec la pandémie ?

La deuxième question est en fait une sous-question de la première, mais suffisamment large pour être considérée comme une question à part entière. Ces deux questions utiliseront des modèles zonaux, tandis que la dernière nécessitera une modélisation ponctuelle.

4.3.1 Q.1 : Existe-t-il un effet COVID-19 significatif sur les accidents de la route au Québec ?

En observant le nombre d'accidents par gravité à la section 3.1, il est clair que l'année 2020, marquée par le début de la pandémie, se distingue des autres. Avec des minimums atteints pour les nombres d'accidents légers et graves, nous avons toutes les raisons de penser que cette année est particulièrement atypique. Pour les accidents mortels, l'année 2020 se démarque beaucoup moins et se confond même avec les autres années.

Évidemment, la diminution du nombre de véhicules sur les routes pendant la pandémie peut expliquer les baisses observées des accidents graves et légers. Cependant, nous avons pris en compte cette exposition dans les modèles BYM. Le nombre d'accidents attendus dans chaque zone (municipalité et MRC) à chaque année est notamment estimé à partir du débit journalier moyen annuel. Cette première question de recherche vise donc à déterminer si, depuis le début de la pandémie, le nombre d'accidents a significativement changé.

Indicateur temporel de la COVID-19

Tout d'abord, nous utiliserons les modèles de base BYM finaux (section 4.1.2) auxquels nous ajouterons simplement un indicateur pour l'année 2020 et les suivantes (2021 et 2022 dans notre cas). Bien que la pandémie ait véritablement débuté à la mi-mars, comme les données sont annuelles, nous nous contenterons d'utiliser 2020 et plus. Nous rappelons que les modèles de bases finaux sont les modèles binomiaux négatifs ajustés au niveau des municipalités, avec l'exposition plus fidèle à la quantité de déplacement et sans inclure la région administrative du Nord-du-Québec.

L'indicateur temporel prend cette forme :

Pre-Covid si l'accident a eu lieu entre 2014 et 2019.

Covid si l'accident a eu lieu entre 2020 et 2022.

En ajoutant simplement cet indicateur temporel comme covariable supplémentaire au modèle, nous obtenons, pour les quatre gravités (léger, grave, mortel et mortel+grave), de

meilleurs DIC qu'auparavant. Notons également que les coefficients significatifs restent les mêmes que ceux présentés au tableau 4.6 avec l'ajout de cet indicateur temporel. Par contre, le coefficient de l'indicateur temporel présenté au tableau 4.9 est très intéressant.

Coefficients Gravité	Modèle 1 Mortel	Modèle 2 Grave	Modèle 3 Grave + Mortel	Modèle 4 Léger
ind_covid	0.120 *	-0.117 *	-0.049 *	-0.201 *
DIC	10609	18613	16959	47224

TABLEAU 4.9 – Coefficients de l'indicateur temporel des années 2020 et suivantes, noté *ind_covid*, ajouté comme covariable dans les modèles de base. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité. Les astérisques indiquent que le coefficient est significatif, c'est-à-dire que l'intervalle de crédibilité à 95% ne contient pas zéro.

Ce qui saute aux yeux est d'abord que le coefficient est significatif dans tous les cas, mais positif pour les accidents mortels, alors qu'il est négatif pour les autres gravités. On interprète donc que dans la période pendant la COVID-19, c'est-à-dire de 2020 à 2022, le nombre d'accidents mortels a augmenté d'un facteur d'environ 12.7% dans chaque municipalité, en gardant les autres variables constantes. On note aussi des diminutions de 11% et de 18% pour les accidents graves et légers respectivement, en gardant les autres variables constantes.

De plus, la combinaison des accidents mortels et graves en une seule catégorie atténue l'effet distinct de chaque type d'accident. Toutefois, étant donné qu'il y a généralement environ cinq fois plus d'accidents graves que d'accidents mortels, le coefficient obtenu est beaucoup plus proche de celui des accidents graves. Puisque notre objectif est de comprendre pourquoi les accidents mortels n'ont pas suivi la même tendance que les autres types d'accidents, nous cesserons de modéliser la somme des accidents graves et mortels.

Indice des interventions liées à la COVID-19

Dans MAMRI et al., 2023, un indice mesurant les interventions non pharmaceutiques au niveau des régions administratives québécoises a été élaboré. Cet indice prend en

compte diverses mesures comme la fermeture d'établissements d'enseignement, d'entreprises non essentielles et de prisons, les couvre-feux, les annulations et suspensions d'événements publics, ainsi que la promotion du télétravail. L'indice est disponible quotidiennement du 1er janvier 2020 au 31 décembre 2022 pour chaque région administrative du Québec. Étant donné que nos données sont annuelles, nous avons fait la moyenne des indices sur toute l'année, aboutissant à trois indices annuels par région administrative. La figure 4.7 illustre les indices annuels pour chaque région administrative. Notons que dans MAMRI et al., 2023, deux indices sont élaborés, mais ces deux indices ont une corrélation de 99% lorsqu'on fait la moyenne annuelle. Par conséquent, la suite se fera uniquement avec ce qu'ils ont appelé l'indice théorique.

FIGURE 4.7 – Indice des interventions liées à la COVID-19 par année et région administrative.

On remarque que toutes les régions administratives suivent des tendances très semblables, probablement car prendre la moyenne annuelle de l'indice atténue les différences. Sinon, on remarque que les interventions gouvernementales étaient fortes en 2020 et 2021, mais ont bien chuté en 2022. Finalement, les valeurs attribuées à cet indice de 2014 à 2019 sont 0.

Nous avons ensuite entraîné des modèles BYM identiques aux modèles de base, mais en ajoutant cet indice comme covariable (et en excluant l'indicateur temporel de 2020 et plus). D'abord, en ce qui concerne le DIC, les modèles qui incluent cet indice sont meilleurs que ceux qui ne l'incluent pas. Cependant, les modèles ajustés avec l'indicateur temporel « 2020 et plus » s'ajustent encore mieux aux données. Le tableau 4.10 présente les coefficients de cet indice pour chaque gravité. À noter que les autres coefficients restent presque identiques à ceux des modèles de base (voir tableau 4.6)

Coefficients	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Gravité	Mortel	Grave	Grave + Mortel	Léger
ind_intervCOVID19	0.344	-0.421 *	-0.144	-0.737 *

TABLEAU 4.10 – Coefficients de l'indicateur des interventions liées à la COVID-19, noté *ind_intervCOVID19*, ajouté comme covariable dans les modèles de base. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité. Les astérisques indiquent que le coefficient est significatif, c'est-à-dire que l'intervalle de crédibilité à 95% ne contient pas zéro.

On remarque les mêmes tendances de coefficients positifs pour le modèle ajusté sur les accidents mortels, et négatifs pour tous les autres. Cependant, le coefficient du modèle ajusté sur les accidents mortels n'est pas significatif, probablement parce que 2022 a un indice très faible, alors que les accidents mortels ont continué d'augmenter. Malgré tout, on peut confirmer l'exclusion de la combinaison des accidents graves et mortels, car encore une fois, les coefficients des modèles d'accidents graves et mortels indiquent des directions différentes.

Modèle de Bernardinelli

Nous avons établi que l'effet temporel post-2020 était significatif, et qu'il est dans un sens opposé pour les accidents mortels. Maintenant, voyons si des modèles spatio-temporels sont en mesure de nous en apprendre davantage.

Commençons par le modèle de Bernardinelli, une extension spatio-temporelle du modèle BYM qui introduit ω, une tendance linéaire globale, et δ_i, appelée tendance différen-

tielle, qui mesure à quel point la tendance temporelle de la région i s'écarte de la tendance globale ω. La formulation générale du modèle est donnée à l'équation 2.37.

Dans ce cadre, la tendance linéaire globale ω est un effet temporel linéaire dépendant des années, où t_j représente les années de 2014 à 2022. Cependant, comme l'année 2020 et les suivantes sont marquées par des particularités liées à la pandémie de COVID-19, nous supposons que la tendance linéaire est différente avant et après 2020.

Par conséquent, nous allons ajuster un modèle de Bernardinelli avec une interaction entre la tendance linéaire globale ω et un indicateur temporel post-2020 (indCovid_j). Ce dernier prend la valeur 0 avant 2020 et 1 à partir de 2020. Cela nous conduit à considérer deux tendances globales distinctes : une avant 2020 et une après. De la même manière, nous introduisons une interaction entre les tendances différentielles δ_i et l'indicateur temporel COVID, permettant à chaque région d'avoir des tendances temporelles spécifiques pour ces deux périodes. Le modèle final prend la forme suivante :

$$\ln(\theta_{ij}) = X_{ij}\beta + b_i + v_i + (\omega : \text{indCovid}_j + \delta_i : \text{indCovid}_j) \cdot t_j,$$

où i est un index des régions spatiales, j, un index temporel, et θ_{ij} le risque relatif de la région i au temps j.

Comme les coefficients des covariables restent pratiquement identiques à ceux présentés dans le tableau 4.6, et les variables significatives sont les mêmes, nous présentons uniquement les tendances globales ω pré et post 2020 au tableau 4.11.

Coefficients	Modèle 1	Modèle 2	Modèle 3
Gravité	Mortel	Grave	Léger
annee :Pre-Covid	-0.018	-0.038 *	-0.016 *
annee :Covid	0.007	-0.031 *	-0.03 *

TABLEAU 4.11 – Coefficients des tendances linéaires globale par municipalités avant *annee :Pre-Covid* et après (inclusivement) 2020 (*annee :Covid*) des modèles de Bernardinelli ajustés. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité. Les astérisques indiquent que le coefficient est significatif, c'est-à-dire que l'intervalle de crédibilité à 95% ne contient pas zéro.

Les DIC des modèles spatio-temporels sont presque identiques à ceux des modèles de base, suggérant une amélioration limitée. En analysant les résultats, nous constatons que les tendances linéaires globales ne sont pas significatives pour les accidents mortels, contrairement aux autres types d'accidents. Il est cependant intéressant de noter que, pour les accidents mortels, les tendances linéaires globales ont des signes opposés avant et après 2020, bien que ces résultats ne soient pas significatifs. Cela indique une possible inversion de tendance, mais on ne peut conclure cela en raison du manque de significativité statistique. Ainsi, cette modélisation ne permet pas de dégager de conclusions robustes quant à l'évolution des accidents mortels.

Modele Knorr-Held

Le modèle Knorr-Held (équation 2.38) permet plus de flexibilité dans la modélisation de l'effet temporel. Contrairement à l'hypothèse de linéarité imposée par le modèle de Bernardinelli, cet effet est ici représenté par deux effets aléatoires distincts modélisant les dynamiques temporelles : un effet non structuré (iid) et l'autre suivant une marche aléatoire du premier ordre. Cette approche permet de mieux capturer des dynamiques temporelles complexes, telles que des variations non linéaires, des changements brusques ou des périodes de stagnation.

Malgré cette flexibilité accrue, les résultats du DIC et de coefficients estimés restent très proches de ceux obtenus avec les modèles précédents. Pour approfondir l'analyse, examinons alors la somme des deux effets aléatoires temporels, $\gamma_j + \phi_j$, à la figure 4.8. Cette somme nous permet de visualiser la dynamique temporelle globale, après avoir contrôlé pour les autres facteurs explicatifs comme les covariables ou l'exposition.

À la différence d'autres approches utilisées dans ce mémoire, où l'année 2020 est explicitement considérée comme un point de rupture (indicateur COVID-19), cette approche est plus descriptive et observatoire. Elle examine l'évolution temporelle sans supposer un moment exact où la pandémie a affecté les résultats.

Nous observons une tendance à la hausse des effets aléatoires pour les accidents mortels au fil des années, tandis que les accidents graves et légers montrent une tendance à

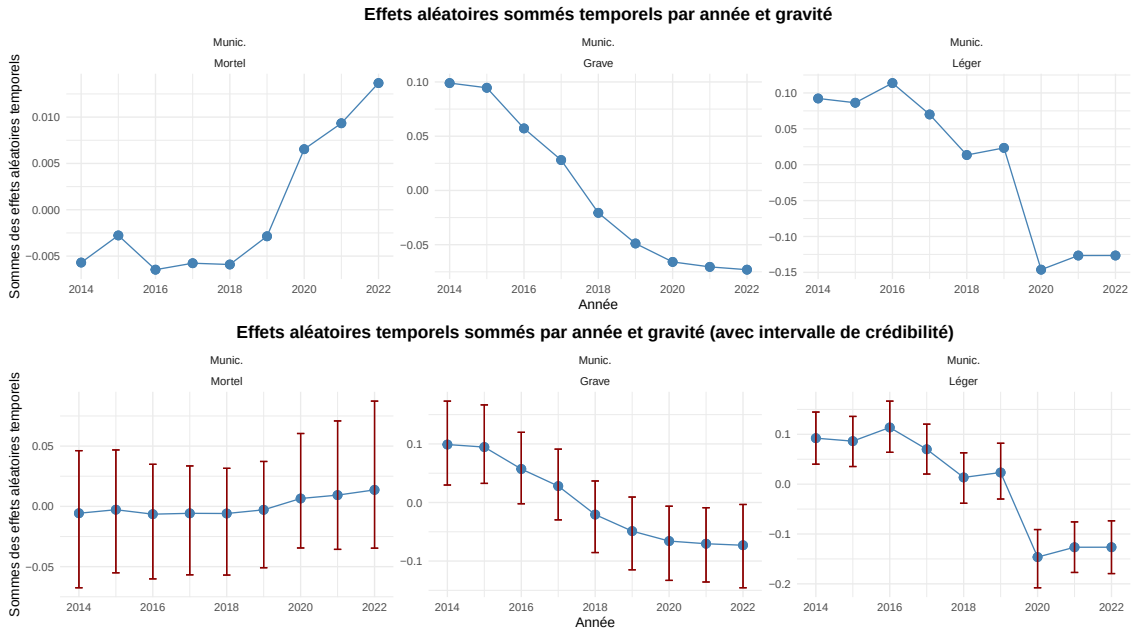

FIGURE 4.8 – Somme des effets aléatoires temporels $\gamma_j + \phi_j$ des modèles Knorr-Held. En-bas, on y ajoute les intervalles de crédibilité. L'analyse est faite au niveau des municipalités.

la baisse. Cela est en soi très intéressant. En effet, les modèles indiquent un changement important des effets temporels en 2020, en particulier pour les accidents mortels et légers.

Cependant, les intervalles de crédibilité de la somme des effets aléatoires temporels individuels incluent fréquemment zéro, du moins pour les accidents mortels. Ainsi, cette analyse ne permet pas de tirer de conclusions solides pour ce type de gravités. En revanche, une tendance significative est observée pour les accidents légers, avant et après 2020. Cela nous permet de confirmer (encore une fois) qu'une modification des tendances temporelles a bien eu lieu, et que la COVID-19 a eu un impact concret sur les accidents légers. Pour les accidents graves, la tendance à la baisse des effets aléatoires est significative si on regarde les deux années extrêmes (2014 et 2022), mais il s'agit d'un phénomène déjà enclenché avant la pandémie.

Finalement, bien que les modèles spatio-temporels plus complexes nous apportent des résultats intéressants, il n'y a pas grand-chose qui soit significatif pour les accidents mortels. Cela est probablement dû à un nombre insuffisant de données par municipalité pour

les accidents graves et mortels. Nous allons donc nous concentrer sur des modèles spatiaux où on segmente les données annuellement, en indiquant simplement l'effet temporel à l'aide d'un indicateur pré et post 2020.

Segmentation mensuelle des données

Une question importante se pose : obtiendrait-on des résultats différents en analysant les données sur une période temporelle différente ? Depuis le début de notre analyse, nous avons segmenté les données par année, mais il est aussi possible de le faire par mois. L'analyse mensuelle permettrait de mieux saisir l'impact de la COVID, en affinant l'indicateur temporel à partir de mars 2020.

Étant donné que nous avons déjà peu de données au niveau municipal en segmentant par année, nous sommes contraints d'effectuer l'analyse mensuelle en utilisant les données segmentées par MRC. Comme nous l'avons vu plus tôt, avec l'utilisation du voisinage MRC, moins de 2% de la variance est expliquée par la structure spatiale pour les modèles avec accidents graves ou mortels. Nous conserverons tout de même la structure spatiale du modèle BYM. De plus, nous avons ajouté un effet aléatoire saisonnier pour chaque mois de l'année. Cet effet aléatoire est spécifié sur la bibliothèque *INLA* , et expliqué dans des documents disponibles sur *inla.r-inla-download.org* 2020.

Pour pouvoir comparer, nous avons entraîné des modèles segmentés par année et par mois avec un indicateur COVID temporel et avec un voisinage utilisant les MRC. Pour les années, l'indicateur reste le même que précédemment, c'est-à-dire 2020 et au-delà. Pour les mois, il commence à partir de mars 2020. Les coefficients obtenus sont présentés au tableau 4.12.

Les coefficients significatifs et les DIC sont tous relativement similaires par niveau de gravité. Les seules exceptions entre les modèles ajustés par année et par mois sont *vitesseAutorisee* pour les accidents légers et *ptsMoyens* pour les accidents graves.

Pour l'indicateur temporel COVID, nous constatons des résultats similaires à ceux observés lors de l'analyse par année, avec des coefficients significatifs positifs pour les accidents mortels et négatifs sinon.

Coefficients Gravité Période tempo.	Modèle 1 Mortel Année	Modèle 2 Mortel Mois	Modèle 3 Grave Année	Modèle 4 Grave Mois	Modèle 5 Léger Année	Modèle 6 Léger Mois
(Intercept)	7.218 *	7.893 *	7.084 *	7.119 *	1.932	1.277
densitéLinéaire_routesMTMD	-0.182 *	-0.185 *	-0.112 *	-0.118 *	-0.045	-0.04
pourcentage_moins5Ans	0.034	0.005	-0.005	-0.013	0.048	0.057
pourcentage_femmes	-0.13 *	-0.138 *	-0.141 *	-0.152 *	-0.032	-0.025
pourcentage_jeunes	-0.039	-0.036	-0.002	-0.003	0.021	0.027
pourcentage_baccalaureatOuPlus	0.015	0.009	0.023	0.026	0.019	0.018
pourcentage_secondaireOuMoins	0.038	0.039	0.025	0.025	0.02	0.016
pourcentage_chomeurs	0.074	0.058	0.081	0.093	0.031	0.034
revenuMedian	-0.163 *	-0.156 *	-0.122 *	-0.128 *	-0.168 *	-0.179 *
ptsMoyens	0.034 *	0.25 *	-0.018	0.093 *	0.026	0.005
vitesseAutorisee	0.002	-0.001	0.005 *	0.01 *	-0.004	0.002 *
ind_covid	0.086 *	0.099 *	-0.154 *	-0.14 *	-0.219 *	-0.239 *

TABLEAU 4.12 – Coefficients des modèles BYM segmentés par année et mois, mais par MRC. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité. Les astérisques indiquent que le coefficient est significatif, c'est-à-dire que l'intervalle de crédibilité à 95% ne contient pas zéro.

Nous avons également repris l'analyse en utilisant l'indice des interventions liées à la COVID-19 (section 4.3.1), moyenné mensuellement. Cependant, ici aussi, cet indice ne s'est pas avéré significatif dans les modèles BYM ajustés sur les accidents mortels.

Conclusion

Nous avons été en mesure de montrer qu'un indicateur temporel de la COVID-19 était significatif, peu importe que la segmentation soit par municipalités ou par MRC, et par année ou par mois. Ce qui est particulièrement intéressant, c'est que cet effet est significatif et positif pour les modèles d'accidents mortels, alors qu'il est significatif mais négatif pour les accidents graves et légers. Par conséquent, nous pouvons répondre à la question initiale par l'affirmative : il existe un effet COVID-19 significatif. Par contre, nous n'avons pas été en mesure de montrer que cet effet était significatif pour un indice des interventions gouvernementales. Ceci laisse penser que ce ne sont pas nécessairement les mesures elles-mêmes qui expliquent le changement observé du nombre d'accidents pendant la COVID-19.

À l'aide de modèles spatio-temporels, nous n'avons pas été en mesure de montrer qu'une tendance temporelle linéaire (sur l'échelle logarithmique) globale changeait si-

gnificativement pour les accidents mortels (ou graves) avant et après 2020. Nous n'avons pas non plus été en mesure de trouver des effets aléatoires temporels significatifs. Les résultats montraient bel et bien que les accidents mortels augmentaient après 2020, et que les accidents grave et léger diminuaient, mais de manière non significative.

Finalement, les modèles plus simples, prenant en compte un voisinage spatial et représentant le temps simplement comme un pré et post 2020 sont ceux que nous privilégierons pour la suite de l'analyse, car leur DIC était souvent meilleur que celui des modèles plus complexes, tout en conservant des résultats très similaires. Malheureusement, en examinant les prédictions de ces modèles sur des cartes du Québec, aucune tendance claire ne se dégage. Par conséquent, pour éviter l'encombrement, nous n'inclurons pas ces cartes.

4.3.2 Q.2 : Quels facteurs expliquent l'augmentation significative des accidents mortels après 2020 ?

Tentons maintenant de trouver ce qui explique cet effet temporel significatif post-2020. Nous avons quelques hypothèses principales pour l'expliquer. La première est une augmentation du nombre d'accidents impliquant des usagers vulnérables. Les accidents impliquant des usagers vulnérables sont plus susceptibles d'engendrer des blessures graves que d'autres types d'accidents. Il est logique de penser qu'une augmentation par exemple de 10% des accidents impliquant des usagers vulnérables puisse mener à une augmentation de plus de 10% des accidents mortels totaux. La seconde hypothèse est que ces accidents mortels sont dus à des conducteurs plus distraits qu'auparavant. La troisième est qu'une sous-population dangereuse est responsable de cet effet COVID-19. Finalement, une augmentation des comportements risqués pourrait l'expliquer.

Modèle logistique

La première analyse repose sur un modèle logistique, qui permet d'intégrer des variables difficilement exploitables dans les modèles zonaux, telles que *indDSF*, *Cause* ou *Genre*. Ce modèle nous permet d'identifier certaines variables ayant évolué entre avant et

ind_Covid	Indicateur si l'accident a lieu après le 12 mars 2020
indUV	Indicateur d'implication d'au moins un usager vulnérable dans l'accident (piéton, vélo ou moto)
ind1Veh	Indicateur d'un seul véhicule impliqué dans l'accident
vitesseAutorisee	Vitesse autorisée là où a eu lieu l'accident
points	Nombre de points moyens d'inaptitude accumulés depuis 2 ans des conducteurs impliqués dans l'accident
indDSF	Indicateur d'un dispositif de sécurité fautif
Cause	Cause principale de l'accident selon le policier. Peut prendre les valeurs suivantes : État du conducteur, Comportement risqué, Défauts mécaniques, Distraction, Environnement routier, Infraction au Code de la route, Autres, Rien à signaler
Genre	Genre d'accident selon le policier. Peut prendre les valeurs suivantes : objet fixe, sans collision, collision avec piéton, collision avec cycliste, collision avec véhicule, collision avec animal, autre

TABLEAU 4.13 – Variables utilisées dans le modèle logistique. ind_Covid est la variable dépendante.

après le début de la pandémie. La variable de réponse prend la valeur de 1 lorsque l'accident est survenu après le 12 mars 2020 (jusqu'au 31 décembre 2022), et 0 sinon. Dans le jeu de données, chaque accident est une rangée. Les variables utilisées dans ce modèle logistique sont présentées au tableau 4.13. À noter que les variables *Cause* et *Genre* sont chacune représentée par plusieurs variables binaires, autant que de valeurs possibles moins 1. Nous avons inclus ces variables pour différentes raisons. Par exemple, *indUV* est utilisée pour examiner si l'implication des usagers vulnérables a changé dans les accidents après l'apparition de la COVID-19. La variable *ind1Veh* est employée pour tenter de corroborer une augmentation des distractions. En effet, nous pensons que les accidents n'impliquant qu'un seul véhicule pourraient être le signe d'une distraction du conducteur, puisque l'accident ne peut être attribué à un autre usager, mais plutôt à une faute du conducteur ou à son environnement. D'autre part, *points*, soit le nombre de points d'inaptitude moyens des conducteurs impliqués dans cet accident, est utilisée pour déterminer si une sous-population de conducteurs dangereux s'est créée avec la pandémie. Finalement, les variables *Cause* et *Genre* peuvent être utilisées pour vérifier plusieurs de ces hypothèses. Notons que là où *Cause* et *Genre* étaient manquantes, nous les avons simplement remplacées respectivement par les catégories « Rien à signaler » et « autre ».

Nous avons ensuite examiné les corrélations entre ces variables. La seule corrélation problématique identifiée (-0.835) concernait les variables *ind1Veh* et *genre_collision avec véhicule*. Étant donné que les variables liées au genre sont plus complexes à intégrer dans

des modèles BYM (en raison du niveau de segmentation requis), nous avons décidé de retirer la variable *ind1Veh* de l'analyse actuelle. Nous allons bien entendu analyser cette variable plus tard.

Des modèles logistiques ont été ajustés séparément pour les accidents mortels, graves et légers. Nous avons également ajusté des modèles avec et sans structure spatiale. L'inclusion d'une structure spatiale implique l'inclusion d'effets aléatoires dans le cadre du modèle BYM, combinant des effets iid et ICAR au niveau des municipalités. Bien que cet ajout améliore légèrement le DIC, l'amélioration reste minime. La proportion de la variance expliquée par la structure spatiale est de 48 % pour les accidents mortels, 13 % pour les accidents graves, et 11 % pour les accidents légers. Nous avons donc décidé de poursuivre avec les modèles incluant la structure spatiale. Le tableau 4.14 montre les coefficients des modèles logistiques selon la gravité.

Coefficients Gravité	Modèle 1 Mortel	Modèle 2 Grave	Modèle 3 Léger
(Intercept)	-0.818	-0.679	-0.367
indUV	0.439 *	0.243 *	0.263 *
vitesseAutorisee	-0.004	-0.005 *	-0.008 *
points	-0.036 *	-0.018 *	-0.03 *
indDSFNon-Respecté	0.139	0.15 *	0.112 *
genre_sans.collision	0.202	0.046	-0.064
genre_collision.avec.piéton	-0.333	-0.269	-0.367
genre_collision.avec.véhicule	0.175	0.016	-0.017
genre_collision.avec.cycliste	0.196	-0.194	-0.208
genre_collision.avec.animal	-0.022	0.181	0.374
genre_objet.fixe	0.426	0.214	0.144
genre_autre	-0.643	0.007	0.141
cause_Comportement.risqué	0.007	-0.015	-0.162
cause_Infraction.au.Code.de.la.route	-0.311	-0.234	-0.144
cause_Distraction	0.313	-0.061	-0.099
cause_État.du.conducteur	0.092	0.049	-0.003
cause_Autres	0.079	0.086	0.355
cause_Environnement.routier	-0.048	-0.042	-0.172
cause_Rien.à.signaler	0.454	0.395	0.277
cause_Défauts.mécaniques	-0.587	-0.178	-0.049

TABLEAU 4.14 – Coefficients des modèles logistiques où la variable de réponse est *ind_covid* avec structure spatiale au niveau municipal, selon la gravité.

On remarque d'abord que les variables de genres et causes ne sont jamais significatives, peu importe la gravité. Les seules variables significatives du modèle pour les accidents graves sont *indUV* et *points*. *indUV*, étant positif, va dans le sens d'une de nos

hypothèses, tandis que *points*, étant négatif, va dans le sens contraire d'une autre. Interprétons les coefficients de ces variables pour bien saisir ce que le modèle signifie :

1. *indUV* : Coefficient positif et significatif pour toutes les gravités d'accidents. Le coefficient de 0.439 est bien plus élevé dans le modèle ajusté sur les accidents mortels. Interprétons ce coefficient particulier.

    a) $\exp(0.439) \approx 1.551$. Ce coefficient positif signifie que, toutes choses étant égales par ailleurs, lorsqu'un usager vulnérable (piéton, cycliste et/ou motocycliste) est impliqué dans un accident mortel, la cote que cet accident soit survenu après le début de la pandémie (post-2020) augmente de 55%. Autrement dit, l'implication d'usagers vulnérables dans un accident est significativement associée à une augmentation de la probabilité que cet accident soit survenu après le début de la pandémie. Cela suggère que les usagers vulnérables pourraient expliquer l'effet temporel COVID positif observé pour les accidents mortels, surtout que le coefficient est près de deux fois plus élevé pour les accidents mortels que pour les accidents graves et légers.

2. *points* : Coefficient négatif et significatif pour toutes les gravités. Sa valeur est de plus en plus négative plus l'accident est grave. Interprétons le coefficient du modèle ajusté sur les accidents mortels.

    a) $\exp(-0.036) = 0.965$. Toutes choses étant égales par ailleurs, la cote qu'un accident ait lieu après le début de la pandémie (post-2020) est multipliée par un facteur de 0.96 lorsque le nombre de points d'inaptitude moyens des conducteurs impliqués dans l'accident augmente de 1. Cela signifie que la cote que l'accident soit post-2020 diminue de 4% lorsque les points d'inaptitudes moyens augmentent de 1. Cela suggère que des conducteurs avec plus de points d'inaptitude n'étaient pas plus dangereux post-2020 et n'expliqueraient donc pas l'effet temporel COVID positif observé.

On remarque que l'indicateur des dispositifs de sécurité mal utilisés, *indDSF*, est positif mais n'est significatif que pour les accidents graves et légers. Cela signifie que les

dispositifs de sécurité étaient étaient associés à une plus grande proportion d'accidents graves et légers pendant la pandémie. Pour ce qui est de l'indicateur d'accidents impliquant un seul véhicule, celui-ci n'est significatif que pour les accidents légers.

Modèles BYM avec nouvelles variables intéressantes

Nous allons maintenant ajuster les modèles BYM utilisés précédemment afin de vérifier si les variables identifiées comme significatives dans les modèles logistiques permettent d'expliquer l'effet temporel de la COVID-19. Nous intégrerons des variables concernant les usagers vulnérables ainsi que des indicateurs pour les accidents n'impliquant qu'un seul véhicule et ceux où un dispositif de sécurité n'a pas été correctement utilisé (bien que ces deux dernières ne soient pas toujours significatives dans les modèles logistiques). De plus, nous approfondirons l'analyse de la variable *points*. Malheureusement, inclure les causes et les genres d'accidents segmente trop les données de comptage, ce qui compromettrait l'analyse. De plus, ces coefficients se sont révélés non significatifs dans les modèles logistiques.

Cela dit, il est particulièrement difficile d'inclure des variables telles que celles des usagers vulnérables, des accidents impliquant un seul véhicule ou le mauvais usage des dispositifs de sécurité sans trop segmenter le jeu de données. Par exemple, la simple inclusion du nombre d'accidents impliquant des usagers vulnérables se traduit souvent par une forte corrélation avec le nombre total d'accidents. Prendre la proportion de ce type d'accidents par rapport au total ne résout pas le problème, car la corrélation persiste.

Nous avons donc décidé de tester ces variables individuellement en segmentant le jeu de données à chaque fois. Autrement dit, au lieu de disposer d'un nombre total d'accidents par région, par gravité et par année, nous aurons un nombre d'accidents par région, par gravité, par année et selon la présence ou l'absence d'une nouvelle variable. Ces trois nouvelles variables sont *indUV*, un indicateur d'usagers vulnérables, *indDSF*, un indicateur de dispositif de sécurité fautif, et *ind1Veh*, un indicateur qu'un seul véhicule est impliqué dans l'accident. Pour chacune de ces variables, un nouveau jeu de données est créé, doublant ainsi le nombre de lignes du jeu de données initial, car le jeu est segmenté une fois

de plus. Enfin, notons que le nombre d'accidents attendus (E_{ijg}), utilisé pour ajuster les modèles BYM, a également été recalibré en fonction des nouveaux jeux de données, en ajustant le nombre d'accidents réels. Cette étape vise à éviter que les accidents impliquant des usagers vulnérables, moins nombreux, ne soient systématiquement sous-estimés par rapport à ceux n'en impliquant pas.

Pour tester si ces variables influencent spécifiquement l'effet temporel lié à la COVID-19, nous introduirons une interaction entre elles et *ind_covid*. Nous ferons également cette analyse pour la variable *ptsMoyens*, qui représente le nombre moyen de points d'inaptitude des conducteurs impliqués dans des accidents dans la région. Le tableau 4.15 présente les coefficients de ces modèles pour les accidents mortels, en les comparant au modèle contenant uniquement l'indicateur temporel initial et les covariables de base, qui ont été traitées à la section 4.3.1.

Coefficients	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Gravité	Mortel	Mortel	Mortel	Mortel	Mortel
VariableÉtudié		indUV	ind_DisposFautif	ind1Veh	Pts moyens
(Intercept)	7.827 *	8.042 *	7.91 *	8.014 *	7.824 *
densitéLinéaire_routesMTMD	-0.031 *	-0.031 *	-0.032 *	-0.031 *	-0.031 *
pourcentage_moins5Ans	0.046	0.045	0.045	0.046	0.046
pourcentage_femmes	-0.138 *	-0.138 *	-0.138 *	-0.138 *	-0.138 *
pourcentage_jeunes	-0.032	-0.032	-0.033	-0.033	-0.031
pourcentage_baccalaureatOuPlus	-0.001	-0.001	-0.002	-0.001	-0.001
pourcentage_secondaireOuMoins	-0.003	-0.003	-0.003	-0.003	-0.004
pourcentage_chomeurs	-0.015	-0.014	-0.015	-0.015	-0.014
revenuMedian	-0.226 *	-0.227 *	-0.224 *	-0.225 *	-0.226 *
ptsMoyens	0.311 *	0.298 *	0.295 *	0.283 *	0.294 *
vitesseAutorisee	0.013 *	0.012 *	0.013 *	0.012 *	0.013 *
ind_covid	0.12 *	0.01	0.116 *	0.059	0.072
indUV		-0.321*			
ind_covid :indUV		0.29 *			
indDSF			0.003		
indDSF :ind_covid			0.031		
ind1Veh				-0.118 *	
ind1Veh :ind_covid				0.132	
ptsMoyens :ind_covid					0.052 *

TABLEAU 4.15 – Coefficients des modèles BYM ajustés sur les accidents mortels. Le tableau compare le modèle de base avec un effet temporel lié à la COVID-19 (à gauche) à des modèles où cet effet est expliqué par différentes covariables. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité. La ligne « VariableÉtudiée » correspond à la variable ayant une interaction avec *ind_covid*.

Tout d'abord, nous remarquons que les coefficients des covariables utilisées précé-

demment ne changent pratiquement pas. De plus, il est important de ne pas comparer les DIC car les jeux de données utilisés pour ajuster les modèles ne sont pas identiques.

Concentrons-nous maintenant sur le coefficient de *ind_covid*. Ce dernier n'est plus significatif lorsqu'on ajoute les interactions avec *indUV*, *ptsMoyens* ou *ind1Veh*, alors que les interactions elles-mêmes sont significatives.

En examinant le modèle dans lequel on étudie *indUV*, on observe que le coefficient d'*indUV* (- 0.321) et son interaction avec *ind_covid* (0.290) sont significatifs et de sens opposé. Interprétons ces coefficients :

— Effet d'*indUV* : Lorsqu'un accident mortel a lieu avant 2020 (*ind_covid* = « Pré-Covid »), le nombre d'accidents mortels impliquant des usagers vulnérables est 27.5% ($e^{-0.321}$) moins fréquent que ceux n'impliquant pas d'usagers vulnérables, toutes choses égales par ailleurs et après contrôle pour l'exposition. Le coefficient négatif d'*indUV* est contre-intuitif, mais peut s'expliquer par le fait que les municipalités ayant un plus grand nombre d'usagers vulnérables sont également celles où des mesures de sécurité plus strictes sont en place pour les protéger.

— Effet d'*indUV :ind_covid* : L'effet d'*indUV* sur le nombre d'accident mortel est significativement différent avant et après 2020, lorsque toutes les autres variables sont constantes. Ainsi, après 2020, le nombre d'accidents impliquant des usagers vulnérables est seulement 3.1% ($e^{-0.321+0.290}$) moins fréquent que ceux n'impliquant pas d'usagers vulnérables, toutes choses égales par ailleurs. Cette interaction atténue donc presque complètement l'effet fixe d'*indUV*.

Ce résultat est particulièrement intéressant et suggère que les accidents impliquant des usagers vulnérables jouent un rôle clé dans l'effet temporel positif post-2020 observé sur les accidents mortels. Le fait que l'effet simple d'*ind_covid* devienne non significatif en présence de cette interaction renforce cette idée.

On observe un schéma similaire avec *ind1Veh*, mais l'interaction entre *ind1Veh* et *ind_covid* n'est pas significative, ce qui limite les conclusions que l'on peut tirer pour ce cas.

En ce qui concerne la variable *ptsMoyens*, il est très intéressant de constater que la moyenne des points d'inaptitude par municipalité pour les conducteurs impliqués dans des accidents mortels post-2020 est significativement plus élevée que celle pré-2020. Cela s'oppose cependant à ce qu'on avait trouvé avec le modèle logistique, probablement en raison de l'agrégation des données. Finalement, *indDSF* était prometteur, mais le peu de données accessibles sur l'utilisation de dispositifs de sécurité ne nous permet pas de conclure que cette variable est significative.

Pour approfondir cette analyse, vérifions les résultats obtenus, mais cette fois-ci avec les modèles ajustés au niveau des MRC plutôt qu'au niveau des municipalités. Le tableau 4.16 présente les coefficients de ces modèles. Les autres covariables sont dans le modèle, mais ne sont pas présentées puisqu'elles sont très similaires à ce qu'on avait auparavant (tableau 4.6).

Coefficients	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Gravité	Mortel	Mortel	Mortel	Mortel	Mortel
VariableÉtudié		indUV	ind_DisposFautif	ind1Veh	Pts moyens
ptsMoyens	0.034 *	0.04 *	0.035 *	0.036 *	0.052 *
ind_covid	0.086 *	-0.012	0.083	0.032	0.17 *
indUV		-0.205 *			
ind_covid : indUV		0.239 *			
indDSF			0.008		
indDSF : ind_covid			0.011		
ind1Veh				-0.071	
ind1Veh : ind_covid				0.109	
ptsMoyens : ind_covid					-0.062

TABLEAU 4.16 – Coefficient des modèles BYM entraînés sur les accidents mortels au niveau des MRC. On exclut les covariables puisque les coefficients sont quasi-identiques à ceux des modèles entraînés sur les municipalités (tableau 4.15). La ligne « VariableÉtudiée » correspond à la variable ayant une interaction avec *ind_covid*.

On arrive à des conclusions similaires avec *indUV*. En revanche, *ptsMoyens* n'a plus du tout le même effet. Dans ce cas, *ind_covid* est encore plus positif dans ce modèle que dans le modèle de base, mais l'interaction avec *ptsMoyens* est désormais négative (mais non significative). Il s'agit des seules variables étudiées dans cette section qui aient encore des coefficients ou interactions significatifs au niveau des MRC.

Nous nous retrouvons avec la variable concernant les points d'inaptitude, qui semble

présenter des résultats contradictoires. À un niveau individuel (modèle logistique), les conducteurs ayant plus de points d'inaptitude étaient impliqués dans moins d'accidents mortels post-2020. Au niveau municipal, une moyenne plus élevée des points d'inaptitude chez les conducteurs impliqués dans des accidents mortels augmente le nombre d'accidents mortels post-2020, ce qui est cohérent avec le modèle de base. Cela suggère que certaines régions seraient plus dangereuses post-2020. Enfin, au niveau des MRC, l'effet de la moyenne des points d'inaptitude perd sa significativité. Nous continuerons d'utiliser la variable des points d'inaptitude, mais sans interaction avec *ind_covid*.

Finalement, deux des variables étudiées se démarquent pour expliquer l'indicateur temporel COVID-19 : l'indicateur d'usagers vulnérables (*indUV*) et celui d'accidents n'impliquant qu'un véhicule (*ind1Veh*). Nous allons entraîner des modèles avec ces variables sur les accidents graves et légers. Encore une fois, les coefficients de base restent les mêmes, nous allons donc simplement montrer ceux des variables avec interaction dans le tableau 4.17.

Coefficients Gravité	Modèle 1 Mortel	Modèle 2 Grave	Modèle 3 Léger	Modèle 4 Mortel	Modèle 5 Grave	Modèle 6 Léger	Modèle 7 Mortel	Modèle 8 Grave	Modèle 9 Léger
VariableÉtudié				indUV	indUV	indUV	ind1Veh	ind1Veh	ind1Veh
ptsMoyens	0.311 *	0.124 *	0.027 *	0.298 *	0.128 *	0.029 *	0.283 *	0.121 *	0.033 *
ind_covid	0.12 *	-0.117 *	-0.201 *	0.01	-0.147 *	-0.212 *	0.059	-0.152 *	-0.191 *
indUV				-0.321 *	-0.267 *	-0.354 *			
ind_covid : indUV				0.29 *	0.08	0.116 *			
ind1Veh							-0.118 *	0.009	0.722 *
ind1Veh : ind_covid							0.132	0.065	0.026

TABLEAU 4.17 – Coefficients des modèles BYM ajustés avec les accidents mortels, graves et légers. Seules certaines variables d'intérêt sont présentées. Nous comparons le modèle où on ajoute simplement l'effet temporel COVID (à gauche) à des modèles où l'on tente d'expliquer cet effet à l'aide de différentes covariables. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité.

Portons surtout notre intérêt sur la variable *ind_covid* qui reste relativement stable pour les accidents graves et légers lorsqu'on ajoute les indicateurs d'usagers vulnérables et d'un seul véhicule. Remarquons également que l'effet de *indUV* est assez semblable avant la COVID (*indUV*) pour tous les types d'accidents, mais qu'après le début de la

pandémie, l'effet d'*indUV* (auquel on ajoute maintenant l'interaction *ind_covid :indUV*) diffère selon la gravité. En revanche, pour ce qui est des accidents n'impliquant qu'un seul véhicule, l'effet était très différent selon la gravité avant la COVID (*ind1Veh*), et ces différences se sont quelque peu atténuées après 2020. Enfin, l'ajout d'interaction de ces nouvelles covariables rend *ind_covid* non-significatif pour les accidents mortels, mais pour les autres gravité d'accidents, ces nouvelles variables influencent peu *ind_covid*. En d'autres termes, ces nouvelles covariables semblent expliquer une partie de l'effet temporel COVID positif des accidents mortels depuis 2020.

Nous souhaitons maintenant intégrer les deux variables, *indUV* et *ind1Veh*, dans un seul modèle afin de vérifier si cela rapproche davantage le coefficient d'*ind_covid* des accidents mortels à ceux des coefficients de l'indicateur temporel COVID pour les accidents légers et graves. De plus, nous voulons confirmer que l'augmentation de l'effet de *ind1Veh* après la COVID n'est pas simplement due à l'augmentation de *indUV*. En effet, les accidents impliquant des usagers vulnérables concernent souvent un seul véhicule. Par exemple, une collision entre un véhicule et un piéton n'implique généralement qu'un seul véhicule. Pour explorer cela plus en détail, nous avons segmenté davantage notre jeu de données, aboutissant à quatre cas :

— *indUV* = 1 et *ind1Veh* = 1 : accidents impliquant des usagers vulnérables et un seul véhicule ;

— *indUV* = 1 et *ind1Veh* = 0 : accidents impliquant des usagers vulnérables et plus d'un véhicule ;

— *indUV* = 0 et *ind1Veh* = 1 : accidents sans usagers vulnérables mais impliquant un seul véhicule ;

— *indUV* = 0 et *ind1Veh* = 0 : accidents sans usagers vulnérables et avec plus d'un véhicule.

D'abord, pour confirmer que l'augmentation de l'effet de *ind1Veh* après la COVID n'est pas simplement due à l'augmentation de *indUV*, nous avons calculé la proportion d'accidents impliquant 1 véhicule et des usagers vulnérable, parmi tous les accidents im-

pliquant 1 véhicule, et obtenu environ 50% pour les accidents graves et mortels. Il se pourrait donc que l'effet d'*ind1Veh* soit effectivement simplement un résultat d'*indUV*. On retrouve d'ailleurs une corrélation de 88% entre le nombre annuel d'accidents mortels par municipalités impliquant des usagers vulnérables et ceux n'impliquant qu'un véhicule. Les coefficients du modèle incluant les deux variables sont présentés au tableau 4.18 où nous le comparons aux modèles avec les variables individuelles. Encore une fois, nous n'y incluons pas les autres coefficients puisqu'ils sont très similaires aux résultats précédents.

Coefficients	Modèle 1	Modèle 2	Modèle 3
Gravité	Mortel	Mortel	Mortel
VariableÉtudié	ind1Veh	indUV	indUV
ind1Veh	-0.118 *		-0.066
ind_covid	0.059	0.01	-0.009
ind1Veh :ind_covid	0.132		0.071
indUV		-0.321 *	-0.269 *
ind_covid :indUV		0.29 *	0.26 *

TABLEAU 4.18 – Coefficients des modèles BYM avec les variables *ind1Veh* et *indUV*, individuellement et combinées. Les modèles incluent toutes les covariables habituelles (non montrées ici). Les trois modèles révèlent des interactions entre la « VariableÉtudiée » et l'indicateur temporel COVID-19. Les autres coefficients ne sont pas inclus dans le tableau. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité.

L'effet global d'*ind_covid* ne se rapproche pas de celui observé pour les accidents graves et mortels, qui sont négatifs. De plus, *ind1Veh* n'est plus significatif, tout comme ses interactions. Nous en concluons donc que cette variable n'était probablement significative que parce qu'elle était corrélée avec *indUV*. Nous l'excluons donc de nos analyses futures et ne gardons que les usagers vulnérables.

Modèles zonaux avec nombre d'accidents impliquant des usagers vulnérables comme variable de réponse

Il ne nous reste maintenant qu'une seule variable parmi nos hypothèses qui permet d'expliquer une partie de l'effet COVID temporel des accidents mortels. Il s'agit de

celle des usagers vulnérables. Nous allons maintenant entraîner un modèle en utilisant le nombre d'accidents par zones impliquant des usagers vulnérables comme variable de réponse. Nous nommerons cette variable *nbAcc_UV*. Nous entraînerons aussi un modèle où la variable de réponse sera le nombre d'accidents par zones excluant ceux impliquant des usagers vulnérables. Le tableau 4.19 présente les coefficients des modèles avec ces deux variables ainsi que le modèle habituel avec *nombreAccidents*, simplement le nombre d'accidents par zones.

Coefficients	Modèle 1	Modèle 2	Modèle 3
Gravité	Mortel	Mortel	Mortel
Variable de réponse	nbAcc	nbAcc_UV	nbAcc-nbAccUV
(Intercept)	7.827 *	10.753 *	5.989 *
densitéLinéaire_routesMTMD	-0.031 *	-0.026	-0.039 *
pourcentage_moins5Ans	0.046	-0.027	0.072 *
pourcentage_femmes	-0.138 *	-0.154 *	-0.14 *
pourcentage_jeunes	-0.032	-0.076 *	-0.028
pourcentage_baccalaureatOuPlus	-0.001	-0.001	-0.002
pourcentage_secondaireOuMoins	-0.003	-0.002	0.005
pourcentage_chomeurs	-0.015	-0.008	-0.025
revenuMedian	-0.226 *	-0.155 *	-0.221 *
ptsMoyens	0.311 *	0.197 *	0.293 *
vitesseAutorisee	0.013 *	-0.015 *	0.036 *
ind_covid	0.12 *	0.228 *	0.059
DIC	10609	4931	8345

TABLEAU 4.19 – Coefficients des modèles BYM avec *nombreAccidents*, *nbAcc_UV* et *nbAcc_AjUV* comme variable de réponse. L'analyse suppose une distribution binomiale négative pour les nombres d'accidents par municipalité.

Encore une fois, nous constatons que la majorité des coefficients sont très similaires. Toutefois, il est particulièrement intéressant de noter que *ind_covid* est encore plus élevé pour les accidents mortels impliquant uniquement des usagers vulnérables, alors qu'il n'est même pas significatif pour les accidents excluant ces usagers. Cela confirme la conclusion précédente : les accidents impliquant des usagers vulnérables sont responsables de l'effet temporel COVID significatif et positif. Cependant, on ne retrouve pas les coefficients négatifs d'*ind_covid* des modèles avec accidents graves et légers.

Conclusion

Nous avons tenté d'expliquer l'effet temporel COVID, significatif et positif pour les accidents mortels, en utilisant plusieurs variables, telles que l'utilisation de dispositifs de sécurité, les points d'inaptitude, les accidents impliquant un seul véhicule et ceux impliquant des usagers vulnérables. Nous avons pu démontrer que l'implication d'usagers vulnérables dans un accident mortel expliquait une part significative de cet effet temporel positif et significatif. Cependant, contrairement aux accidents graves et légers, cet effet n'est pas négatif et significatif.

4.3.3 Q.3 : La localisation des accidents sur l'île de Montréal a-t-elle changée avec la pandémie ?

Bien entendu, comme nous répondrons à cette question à l'aide de modèles ponctuels et que les données ponctuelles ne sont disponibles que sur l'île de Montréal, les conclusions ne s'appliqueront qu'à cette région. Il est en effet difficile d'extrapoler des résultats issus de la plus grande ville de la province, qui représente une infime partie du territoire mais près du quart de la population, au reste de la province.

Étant donné que les modèles LGCP peuvent rapidement devenir complexes, nous commencerons par évaluer les coefficients de modèles sans inclure de processus spatial. Il s'agit des mêmes modèles que ceux de la section 4.2.3. Notons également que nous continuerons avec les mêmes périodes que celles des modèles ponctuels de base ponctuels.

Modèles avec covariables uniquement

Regardons simplement les coefficients des modèles de la section 4.2.3. Commençons avec les coefficients des modèles ajustés avec des accidents mortels, présentés au tableau 4.20

Analysons les coefficients plus en profondeur :

1. *revenuMedian* : Significatif pour tous les modèles sauf après mars 2020. Dans les

Coefficients Période	Modèle 1 2012-21 (Complète)	Modèle 2 2012-19 (Pré-COVID)	Modèle 3 2018-19 (pré-COVID)	Modèle 4 2020-21 (COVID)
Intercept	-2.314 ∗	-2.134 ∗	-4.256 ∗	-6.693 ∗
feuCircul_Stop	0.012 ∗	0.012 ∗	0.016 ∗	0.011
hopitaux	-0.007	-0.011	-0.037	-0.001
revenuMedian	-0.011 ∗	-0.015 ∗	-0.021 ∗	0.013
pourcentage_chomeurs	0.129 ∗	0.068	0.059	0.515 ∗
densitéPopulation	0.004 ∗	0.004 ∗	0.004	0.004
routes	0.361 ∗	0.344 ∗	0.505 ∗	0.461 ∗
pourcentage_zoneResidentielle	0.014 ∗	0.017 ∗	0.026 ∗	0.008
pourcentage_zoneIndusComm	0.007	0.009	0.008	0.003
pourcentage_jeunes	-0.01	0.01	0.019	-0.14 ∗
pourcentage_ainés	0.017	0.01	0.023	0.05 ∗

TABLEAU 4.20 – Coefficients des modèles LGCP ajustés sur les données d'accidents mortels à Montréal, en fonction des covariables uniquement. Les modèles diffèrent selon les périodes temporelles considérées.

autres modèles, la variable a un coefficient négatif. Voici son interprétation en prenant le coefficient -0.011. Pour chaque augmentation de 1000$ du revenu médian dans le secteur de recensement (SR), l'intensité de 2012 à 2021 des accidents mortels (le nombre attendus d'accidents mortels par kilomètre carré) diminue d'un facteur de 1.1% $\left(\exp(-0.011) \approx 0.989\right)$, en gardant toutes les autres variables constantes. Remarquons que le revenu médian n'est plus significatif pendant la pandémie, mais il change de signe.

2. *pourcentage_chomeurs* : Il s'agit probablement du résultat le plus intéressant de ces modèles. Ce coefficient est significatif pour les modèles entraînés sur les données COVID et complète (2012 à 2021). Remarquons cependant à quel point le coefficient du modèle entraîné sur les données COVID, de 0.515, est élevé par rapport aux autres. Pour chaque augmentation de 1% des chômeurs dans le SR, l'intensité d'accidents mortels augmente d'un facteur de 67.4% $(\exp(0.515))$ pendant la pandémie, en gardant toutes les autres variables constantes. Ceci pourrait simplement être dû à un taux de chômage élevé en 2020 à Montréal (11.4% selon Statistique Canada), mais ce taux était déjà revenu à la normal en 2021.

3. D'autres résultats intéressants concernent les coefficients de *pourcentage_jeunes* et *pourcentage_ainés* :

a) *pourcentage_jeunes* : une augmentation de 1% de jeunes (âgés entre 15 et 24 ans) dans le SR entraîne une diminution de l'intensité des accidents mortels de 13.1% ($\exp(-0.140)$) pendant la COVID, en gardant toutes les autres variables constantes. Ce résultat indique que l'intensité des accidents diminue dans les zones où résident davantage de jeunes, ce qui est surprenant et contredit en partie l'idée répandue selon laquelle les jeunes sont plus à risque sur la route. Il est important de préciser que ce modèle ne relie pas directement l'intensité des accidents à l'âge des conducteurs. Par ailleurs, ce coefficient est significatif que dans le modèle spécifique à la période COVID.

b) *pourcentage_ainés* : une augmentation de 1% de personnes âgées (âgées de 65 ans et plus) dans le SR entraîne une augmentation de l'intensité des accidents mortels de 5.1% ($\exp(0.05)$) pendant la période COVID, en gardant toutes les autres variables constantes. Ce résultat s'aligne avec l'hypothèse selon laquelle les conducteurs plus âgés peuvent être davantage à risque sur la route, mais on ne peut tirer cette conclusion avec notre modèle car il ne relie pas directement l'intensité des accidents à l'âge des conducteurs, mais plutôt à celui des résidents dans le SR. À noter que ce coefficient est significatif uniquement dans le modèle COVID.

4. *densitéPopulation* : Significatif pour les modèles avec plus de données (complet et pré-COVID) et non-significatif pour les modèles avec moins de données (pré et COVID de même durée). Probablement non-significatif car il n'y a pas assez de données. La densité de population peut être interprétée comme un proxy du trafic.

5. *pourcentage_zoneRésidentielle* : Significatif pour tous les modèles sauf le modèle entraîné sur les données COVID. Avant la pandémie, l'intensité d'accidents mortels était plus élevée dans une zone plus résidentielle.

Les changements dans la significativité des coefficients sont surtout dus au modèle entraîné sur les données COVID. On a en effet remarqué que l'intensité d'accidents depuis la COVID-19 n'est plus expliquée par le revenu médian, ou le fait de survenir en

zone résidentielle. Par contre, l'intensité augmentait avec le pourcentage de chômeurs aux alentours. On semble donc avoir une plus grande intensité d'accidents mortels dans des zones moins résidentielles, plus pauvres et avec plus de chômeurs. Peut-être est-ce simplement parce qu'il y a moins d'accidents au centre-ville, ou autour du centre-ville puisqu'il y avait beaucoup moins de déplacements dans ces régions depuis la pandémie.

Regardons maintenant les mêmes modèles, mais entraînés sur les accidents graves. Les coefficients sont présentés au tableau 4.21. Tout d'abord, on remarque que même pour

Coefficients Période	Modèle 1 2012-21 (Complète)	Modèle 2 2012-19 (Pré-COVID)	Modèle 3 2018-19 (pré-COVID)	Modèle 4 2020-21 (COVID)
Intercept	0.692 *	0.512 *	-1.229 *	-1.136 *
feuCircul_Stop	0.01 *	0.01 *	0.009 *	0.004
hopitaux	0.001	0	-0.004	0.007
revenuMedian	-0.018 *	-0.019 *	-0.015 *	-0.014 *
pourcentage_chomeurs	0.109 *	0.096 *	0.106 *	0.21 *
densitéPopulation	0.004 *	0.004 *	0.005 *	0.005 *
routes	0.241 *	0.251 *	0.318 *	0.178 *
pourcentage_zoneResidentielle	0.012 *	0.012 *	0.012 *	0.011 *
pourcentage_zoneIndusComm	0.005 *	0.005 *	0.007	0.004
pourcentage_jeunes	-0.006	0.002	-0.027	-0.077 *
pourcentage_ainés	0.004	0.005	0.003	0.003

TABLEAU 4.21 – Coefficients des modèles LGCP ajustés sur les données d'accidents graves à Montréal, en fonction des covariables uniquement. Les modèles diffèrent par les périodes temporelles considérées.

les modèles entraînés avec les accidents graves, la variable *pourcentage_chomeurs* est significative, mais son amplitude dans le modèle entraîné sur les données COVID est sur une échelle comparable à celle des autres périodes temporelles. Les variables montrant des changements intéressants sont plutôt *feuCircul_Stop* et *routes* qui ne sont plus significatifs alors qu'ils étaient significatifs et positifs auparavant. La proximité de feu de circulation, de panneaux d'arrêts et de routes importantes n'a plus d'effet sur l'intensité des accidents graves pendant la COVID. Cela pourrait suggérer que ces accidents se sont déplacés vers des zones où les routes sont moins achalandées ou moins importantes. Notons que nous ne prêtons pas trop attention à la variable *pourcentage_zoneIndusComm* puisqu'elle semble non-significative en période pré et COVID de même durée en raison d'un manque de données pour les modèles ne prenant que deux années en compte.

Finalement, mettons ces résultats sur une carte et visualisons les log-intensités des modèles pré et pendant la pandémie, mais pour des temps égaux. Ces visualisations sont à la figure 4.9.

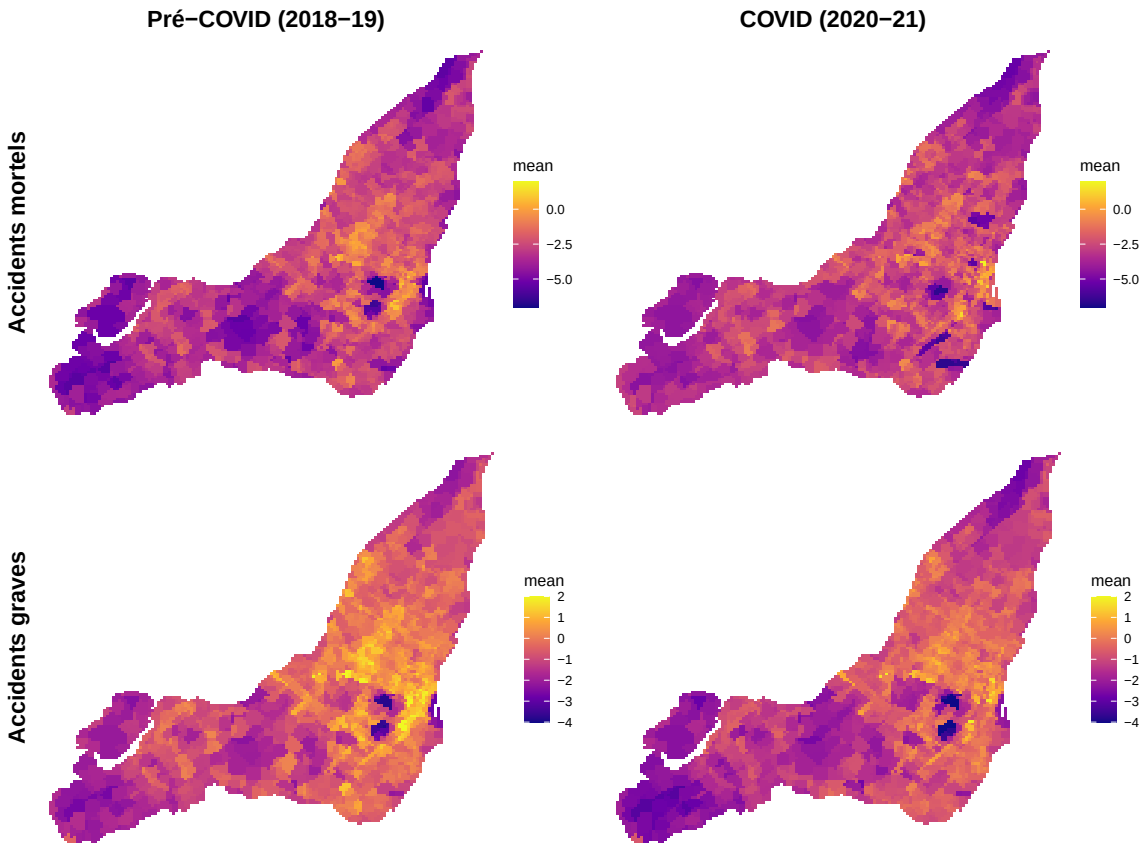

FIGURE 4.9 – Cartes de Montréal montrant les prédictions des modèles LGCP basés uniquement sur les covariables. Les prédictions sont exprimées sur l'échelle de la log-intensité. Les modèles sont entraînés séparément pour les accidents mortels et graves, ainsi que pour deux périodes : pré-COVID (13 mars 2018 - 31 décembre 2019) et COVID (13 mars 2020 - 31 décembre 2021). Les échelles sont constantes selon le type de gravité.

On observe que le centre-ville et le centre de l'île présentent une log-intensité pendant la COVID-19 moins élevée pour les accidents mortels et graves. Cependant, une différence apparaît aux extrémités de l'île : la log-intensité des accidents graves ne semble pas beaucoup varier avant et après la pandémie, tandis que la log-intensité des accidents mortels aux extrémités de l'île est bien plus élevée pendant la pandémie qu'avant. Les accidents mortels semblent être plus uniformément répartis sur l'ensemble de l'île.

Modèles avec processus spatial uniquement

Nous allons maintenant observer les processus gaussiens spatiaux qu'on obtient avec des modèles LGCP où l'on inclut simplement ce processus et bien sûr, l'ordonnée à l'origine. Les modèles LGCP se sont avérés très difficiles à ajuster avec toutes les covariables et le processus spatial, c'est pourquoi nous y allons une étape à la fois. Nous allons simplement présenter les modèles des périodes pré et pendant la COVID-19 de durée égale.

La distribution de la portée ρ et de l'écart-type σ utilisés pour ajuster les modèles LGCP est présenté aux équations 4.3 et 4.4. Ces hyperparamètres jouent un rôle crucial dans la modélisation. Une mauvaise spécification peut entraîner des modèles où le processus spatial reste très proche de 0, n'apportant ainsi aucune information pertinente, ou peut empêcher la convergence des modèles. On rappelle que les hyperparamètres sont définis en fixant les distributions *a priori* à l'aide de l'équation 2.20. Pour les accidents mortels, nous avons utilisé fixé les distributions *a priori* comme suit :

$$P(\rho < 5) = 0.01, P(\sigma > 1) = 0.01, \tag{4.3}$$

alors que pour les accidents graves, nous avons utilisé :

$$P(\rho < 0.1) = 0.01, P(\sigma > 0.5) = 0.01. \tag{4.4}$$

Les unités de mesure sont en kilomètres pour la portée, et sur l'échelle de l'intensité des accidents pour l'écart-type. Plusieurs hyperparamètres ont été testés dans ces modèles, et ceux utilisés semblent logiques. En effet, la portée définit simplement la distance au-delà de laquelle la corrélation entre deux accidents devient négligeable. Ainsi, il est quasiment certain que la portée des accidents mortels est supérieure à 5 km, tandis que celle des accidents graves se limite à environ 100 mètres. On suppose ici que les facteurs locaux influencent davantage les accidents graves sur des distances plus courtes. Par ailleurs, comme les intensités d'accidents sont généralement inférieures à 0.5, il est logique de fixer des écart-types qui seront presque toujours inférieurs à 1 ou 0.5. Plusieurs autres hyperparamètres ont été essayé, avec des résultats assez similaires.

Nous avons cherché à utiliser des hyperparamètres constants au-travers d'une même gravité et permettant de modéliser un processus spatial lisse. La recherche d'hyperparamètres permettant aux modèles de converger s'est avérée relativement simple pour ajuster les différents jeux de données, à l'exception des accidents graves pendant la COVID-19, pour lesquels il a été impossible de trouver des paires d'hyperparamètres permettant au modèle de converger. Nous ne sommes pas en mesure d'expliquer d'où provient cette difficulté, mais elle sera récurrente pour tous les modèles graves entraînés sur les données COVID.

Au tableau 4.22, nous comparons les DIC des modèles avec processus spatial uniquement avec ceux des modèles avec covariables uniquement. Bien que les DIC des modèles

Gravité	Période	Effets	DIC
Mortel	2018-19 (Pré-COVID)	Covariables seules	-314
Mortel	**2018-19 (Pré-COVID)**	**Proc. Gaussien seul**	**-356**
Mortel	2020-21 (COVID)	Covariables seules	-306
Mortel	**2020-21 (COVID)**	**Proc. Gaussien seul**	**-359**
Grave	**2018-19 (Pré-COVID)**	**Covariables seules**	**-2080**
Grave	2018-19 (Pré-COVID)	Proc. Gaussien seul	-2065
Grave	2020-21 (COVID)	Covariables seules	-1695
Grave	**2020-21 (COVID)**	**Proc. Gaussien seul**	**-1861**

TABLEAU 4.22 – Comparaison des DIC entre les modèles LGCP basés uniquement sur les covariables et ceux basés uniquement sur le processus spatial.

soient à peu près sur la même échelle, dans trois cas sur quatre, les modèles avec processus gaussien uniquement sont meilleurs selon le DIC. Regardons, à la figure 4.10, les cartes de Montréal obtenues avec les prédictions des modèles avec processus spatial uniquement pour voir si cela concorde avec les résultats obtenus à partir des modèles basés exclusivement sur les covariables. Les tendances observées ici semblent similaires à celles issues des modèles basés uniquement sur les covariables. Toutefois, cette constatation reste purement visuelle, et il est difficile d'en tirer des conclusions définitives en raison des différences significatives dans le niveau de détail des projections. Notamment, on observe une intensité des accidents mortels mieux répartie pendant la COVID-19, avec des intensités moins faibles aux extrémités et moins élevées au centre de l'île. De plus, pour les accidents graves, on remarque encore ici une intensité des accidents graves moins

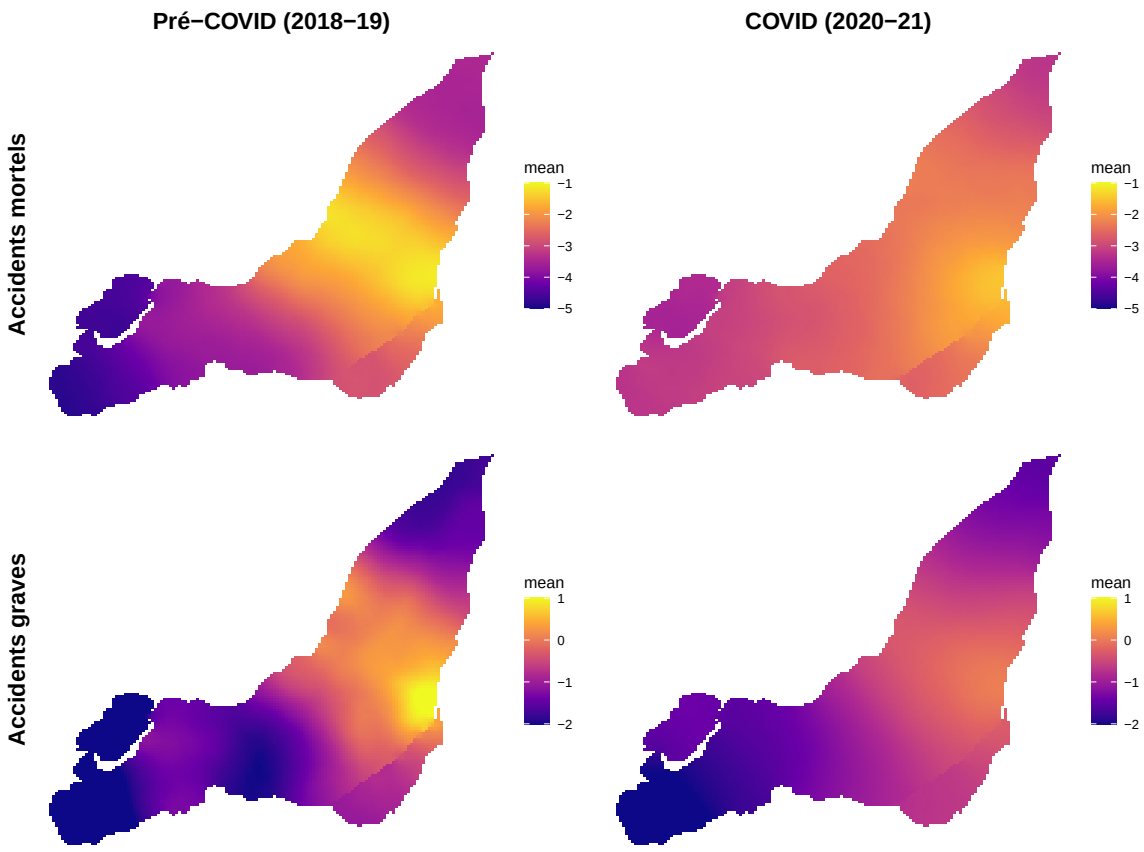

FIGURE 4.10 – Cartes de Montréal montrant les prédictions des modèles LGCP basés uniquement sur le processus spatial (et l'ordonnée à l'origine). Les prédictions sont exprimées sur l'échelle de la log-intensité. Les modèles sont entraînés séparément pour les accidents mortels et graves, ainsi que pour deux périodes : pré-COVID (13 mars 2018 - 31 décembre 2019) et COVID (13 mars 2020 - 31 décembre 2021). Les échelles sont constantes selon le type de gravité.

forte au centre-ville depuis la COVID-19.

Modèles avec covariables et processus spatial

Nous abordons à présent des modèles plus complets, permettant d'intégrer un processus spatial tout en contrôlant les covariables. Ces modèles ont été ajustés avec les mêmes distributions d'hyperparamètres que celles présentées à la section 4.3.3

Concernant le DIC, les modèles incluant à la fois les covariables et le processus spatial présentent des DIC plus élevés, ce qui indique un ajustement moins performant aux don-

nées par rapport aux modèles utilisant uniquement les covariables ou le processus spatial. Cela suggère que la combinaison des covariables et du processus spatial rend le modèle trop complexe. Malgré tout, ces DIC restent sur une échelle comparable.

La figure 4.11 montre les prédictions du processus spatial de ces modèles. Il s'agit d'un processus spatial qui explique les variations non capturées par les covariables. Notons que les covariables sont presque identiques avec ou sans le processus spatial. On

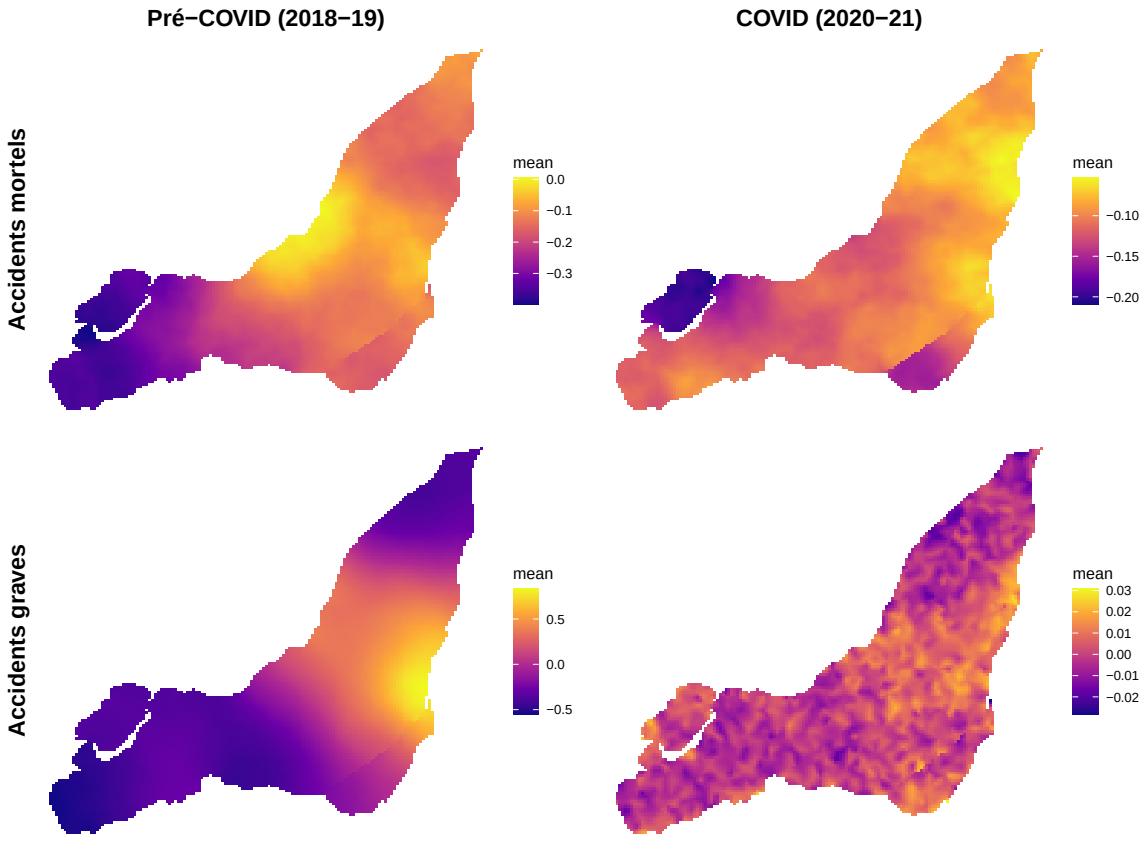

FIGURE 4.11 – Cartes de Montréal illustrant les processus spatiaux des modèles LGCP combinant processus spatial et covariables. Les processus spatiaux $Z(s)$ sont représentés sur l'échelle de la log-intensité. Les modèles sont entraînés séparément pour les accidents mortels et graves, et pour deux périodes : pré-COVID (13 mars 2018 - 31 décembre 2019) et COVID (13 mars 2020 - 31 décembre 2021).

observe que, pour les accidents mortels et graves, le processus spatial pendant la COVID-19 est très proche de zéro, indiquant une faible intensité résiduelle. Avant la pandémie, cette intensité était légèrement plus élevée, bien que toujours faible. Cela suggère que

les covariables expliquent mieux les effets spatiaux pendant la pandémie qu'auparavant. Cependant, dans les deux périodes, les covariables semblent jouer un rôle nettement plus déterminant que le processus spatial sous-jacent. Enfin, on note que ces covariables n'expliquent pas entièrement la forte intensité des accidents mortels observée avant la COVID-19 dans le centre-nord de l'île, ni celle des accidents graves au centre-ville, où les processus spatiaux semblent mieux capturer ces dynamiques locales.

Afin de vérifier si la grande intensité observée au centre-nord de l'île de 2018 à 2019 reflète une tendance plus ancienne, nous avons entraîné un modèle sur l'ensemble de la période pré-COVID disponible, soit du début de 2012 au 13 mars 2020, pour les accidents mortels. Comme on peut le voir à la figure 4.12, on retrouve plutôt une intensité plus élevée vers le centre-ville. L'intensité au centre-nord de l'île de 2018 à 2019 semble donc être circonstancielle.

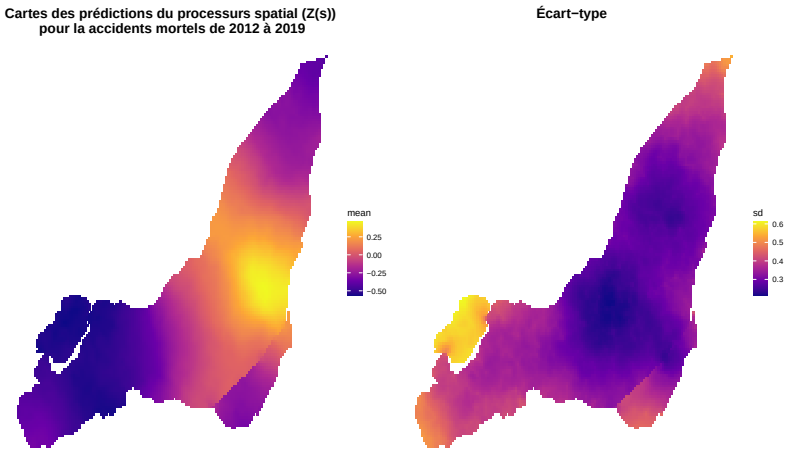

FIGURE 4.12 – Cartes de Montréal illustrant le processus spatial et son écart-type des modèles LGCP combinant processus spatial et covariables. Les processus spatiaux $Z(s)$ sont représentés sur l'échelle de la log-intensité. Le modèle est entraîné sur les accidents mortels entre 2012 et 2019.

Il est néanmoins important de souligner que les paramètres de portée et d'écart-type du processus spatial présentent des écarts types très élevés (parfois aussi grands que les paramètres eux-mêmes). On remarque d'ailleurs que l'écart-type du processus spatial sur l'échelle de la log-intensité à la figure 4.12, a souvent une amplitude presque aussi grande que l'intensité elle-même. Cela peut indiquer que les distributions *a priori* n'étaient pas

adéquates. Nous en avons essayé plusieurs autres, mais n'avons pas obtenu de meilleurs résultats. Il est donc difficile de conclure quoi que ce soit à partir des processus spatiaux obtenus.

Malgré tout, en comparant les coefficients des modèles incluant un processus spatial avec ceux des modèles ne considérant que les covariables, on constate très peu de différences notables. Cela suggère que les effets spatiaux sont probablement faibles ou difficiles à identifier dans ces modèles, ce qui est en accord avec la variabilité importante observée.

Modèle multivarié - Modélisation de la différence entre les périodes d'avant et pendant la COVID-19

Nous avons aussi essayé d'ajuster les modèles décrits à l'équation 2.25. On cherchait alors à modéliser l'intensité des périodes pré-COVID-19 et COVID-19 pour chaque gravité, en espérant cette fois-ci modéliser des processus spatiaux moins variables. Pour ce faire, nous avons modélisé deux processus spatiaux par gravité, un décrivant ce qu'avait en commun les intensités des périodes pré et pendant la pandémie, et l'autre décrivant plutôt ce qui les différenciait. Ces modèles sont difficiles à ajuster puisqu'ils demandent des distributions *a priori* d'hyperparamètres pour les deux processus spatiaux. Malheureusement, encore une fois, les paramètres des processus spatiaux présentaient des écart-types très élevés.

Conclusion

Finalement, l'analyse ponctuelle des accidents s'est révélée très difficile et n'a pas donné de résultats clairs. Les résultats les plus intéressants proviennent de l'interprétation des covariables, qui indiquent une plus grande intensité d'accidents mortels dans des zones moins résidentielles, plus pauvres, où moins de jeunes résident et avec plus de chômeurs pendant la COVID-19. Malheureusement, mis à part la confirmation sommaire de certaines tendances issues des modèles à covariables uniquement, les modèles avec

processus spatiaux ne nous ont pas permis de tirer de conclusions significatives. Nous attribuons cela, entre autres, au fait que ces modèles sont difficiles à ajuster et dépendent beaucoup d'hyperparamètres qui sont difficiles à trouver, mais aussi au fait qu'ils sont très sensibles aux points lorsque nous avons peu de données, ce qui est le cas avec les accidents mortels et même graves. Même si nous avions plus de données, ces modèles deviennent très rapidement longs à entraîner. Malgré tout cela, le domaine de la modélisation ponctuelle, et particulièrement les modèles LGCP sur des bibliothèques comme *inlabru*, connaît un essor très important depuis quelques années. Il est permis de croire que cette analyse deviendra bien plus intéressante dans un avenir proche.

Conclusion

Revenons d'abord sur ce qui a motivé ce projet : l'étude de l'évolution de la sécurité routière au Québec avant et après la pandémie de COVID-19. Une observation majeure a été l'augmentation significative des accidents mortels, parallèlement à une diminution des accidents graves et légers. Cette tendance a attiré l'attention, car elle suggérait que l'impact de la pandémie sur les comportements de conduite et les conditions de circulation avait pu être plus marqué pour les accidents les plus graves. Après avoir étudié les données, nous avons confirmé que l'augmentation des accidents mortels était statistiquement significative, tout comme la diminution des accidents graves et légers, une tendance qui se reflétait clairement dans les données aux échelles municipale et régionale (MRC).

Pour mieux comprendre l'évolution spatio-temporelle des accidents, nous avons utilisé des modèles intégrant à la fois des effets spatiaux et temporels, tels que le modèle de Bernardinelli et celui de Knorr-Held. Bien que ces modèles aient permis d'identifier certaines tendances intéressantes, les résultats pour les accidents mortels ne se sont pas révélés significatifs dans la majorité des cas. Cela est probablement dû au nombre insuffisant de données par municipalité pour ce type d'accident, rendant difficile l'identification d'un signal robuste dans les variations spatio-temporelles. En conséquence, nous avons privilégié une approche plus simple mais plus robuste, en segmentant les données par année et en représentant l'effet temporel à l'aide d'un indicateur pré- et post-2020. Cette simplification a permis de mieux capter l'impact global de la pandémie sur la sécurité routière, ce qui a permis de capter l'impact global de la pandémie tout en limitant un surajustement des modèles.

L'une des contributions principales de cette étude a été de montrer qu'une partie importante de cette augmentation des accidents mortels était portée par les accidents impliquant des usagers vulnérables, tels que les piétons, les cyclistes et les motocyclistes. Cette observation a permis de souligner l'importance d'une meilleure prise en charge de ces catégories d'usagers, en particulier dans un contexte où des mesures sanitaires ont modifié la mobilité des citoyens. En effet, la pandémie a pu entraîner des changements dans les flux de circulation, avec potentiellement plus de déplacements à pied ou à vélo, en raison des restrictions de mobilité et du télétravail.

Pour tester précisément l'effet des usagers vulnérables, nous avons intégré une interaction entre l'indicateur temporel COVID-19 (pré-/post-2020) et la présence d'usagers vulnérables dans nos modèles. Cette approche a permis d'évaluer si la hausse des accidents mortels après 2020 était plus marquée pour les accidents impliquant des usagers vulnérables que pour ceux n'en impliquant pas. Les résultats ont montré que cette interaction était statistiquement significative, suggérant que la pandémie a eu un impact différent selon le type d'usagers impliqués. De plus, pour mieux isoler cet effet, nous avons construit des modèles séparés où la variable réponse était le nombre d'accidents mortels impliquant des usagers vulnérables d'une part, et ceux n'en impliquant pas d'autre part. Cette segmentation a confirmé que l'augmentation post-2020 était principalement portée par les accidents impliquant des usagers vulnérables, tandis que ceux n'en impliquant pas n'ont pas montré de variation significative.

Finalement, nous avons pu observer des changements dans l'intensité des processus spatiaux pendant la pandémie, avec notamment une répartition plus homogène des accidents sur l'île de Montréal. Cette tendance suggère que les modifications des flux de circulation et des habitudes de déplacement ont eu un impact sur la localisation des accidents mortels. Il est possible que ces changements reflètent une adaptation des comportements de mobilité, affectant ainsi la répartition géographique des accidents. Cependant, ce sont surtout les covariables intégrées aux modèles qui ont permis d'obtenir des conclusions significatives, mettant en évidence des facteurs socio-économiques associés à une plus grande intensité des accidents mortels. L'analyse a révélé que les zones les plus touchées

par les accidents mortels pendant la pandémie étaient caractérisées par une proportion plus faible de jeunes résidents, un taux de chômage accru et des zones moins résidentielles. Ces résultats soulignent l'importance de considérer les effets socio-économiques dans l'analyse des accidents de la route, en particulier lors de chocs externes majeurs comme la pandémie, qui ont pu modifier les profils de risque des différentes zones urbaines.

Cependant, plusieurs limitations et difficultés ont émergé au cours de cette analyse. D'abord, il est important de noter que les choix des modélisations n'étaient peut-être pas les plus judicieux pour les données sous la main. En effet, il était très difficile d'étudier les accidents en faisant des comptes par régions puisque cela demandait une segmentation des données toujours plus fine. Cette segmentation a conduit à des résultats parfois imprécis, en particulier lorsque les données étaient trop peu nombreuses, ce qui a limité la robustesse des modèles à des échelles fines.

Les deux principaux modèles utilisés (BYM et LGCP) utilisés dans ce projet permettent d'analyser les facteurs démographiques et socio-économiques, mais sont moins adaptés à l'analyse des différences entre différents types d'accidents. Pour les modèles BYM, cela est principalement lié à un manque de données. Dans le cas des accidents mortels, ce manque s'explique par le fait qu'il n'y a heureusement pas suffisamment d'accidents mortels pour constituer un échantillon représentatif lorsqu'on segmente à des niveaux un peu plus fins que celui des municipalités ou de l'échelle annuelle. En ce qui concerne les modèles LGCP, il s'agit d'un domaine avec beaucoup de recherches ces dernières années, et il deviendra peut-être plus simple d'analyser et de modéliser les marques ou valeurs associées aux points spatiaux, mais ce n'est malheureusement pas encore le cas, ce qui a limité l'utilité de ces modèles dans le projet. Il aurait peut-être été préférable d'approfondir l'un des deux types de modélisation (BYM ou LGCP), mais des difficultés d'accès aux données et des problèmes de confidentialité nous ont contraints à consacrer plus de temps que prévu aux modèles ponctuels.

Par ailleurs, il est important de souligner que les données de 2023, récemment disponibles, confirment les tendances observées entre 2020 et 2022, à savoir une hausse des

décès et une diminution des accidents graves et légers par rapport à la moyenne des cinq années précédentes. L'inclusion de ces nouvelles données pourrait permettre de surmonter certaines limitations liées au faible nombre d'accidents observés. Malheureusement, nous n'avons pas eu accès à ces données.

Malgré tout, nous trouvons les résultats obtenus intéressants, prometteurs et constituent une bonne entrée en matière dans ce projet plus large de sécurité routière visant éventuellement à formuler des recommandations pour des interventions de sécurité routière dans les années à venir. En effet, l'extraction et le prétraitement de données provenant de différentes sources externes, des voisinages de municipalités ajustés pour que des municipalités reliés par un pont soient considérées voisines par contiguïté, et l'exposition calculée à partir de débit journaliers moyens annuels sont tous des éléments qui pourront être repris et adaptés dans d'autres projets.

En conclusion, cette étude met en lumière des tendances préoccupantes concernant l'augmentation du nombre d'accidents mortels au Québec depuis l'arrivée de la COVID-19, tout en soulignant le rôle significatif des usagers vulnérables dans cette dynamique. Les résultats obtenus offrent des pistes précieuses pour les décideurs et les intervenants dans le domaine de la sécurité routière. À l'avenir, il sera crucial de continuer à analyser des données sur les accidents, en intégrant les nouvelles tendances et en adaptant les modèles de prévision. Les codes utilisés dans ce mémoire sont disponibles sur GitHub au lien suivant `https://github.com/edgarLan/SAAQ_SR`.

Bibliographie

ABDOU, M (2023), *A big change is happening in Canada's urban centers – ushering a new era of doing business*, https://bdl-lde.ca/publications/a-big-change-is-happening-in-canadas-urban-centers-ushering-a-new-era-of-doing-business/, Récupéré le 15-04-2024.

ADAM, C (2020), *Saanich sees spike in speeders as roads empty due to COVID-19*, https://vancouverisland.ctvnews.ca/saanich-sees-spike-in-speeders-as-roads-empty-due-to-covid-19-1.4886785, Récupéré le 08-04-2024.

ADANU, E, OKAFOR, S, PENMETSA, P et JONES, S (2022), « Understanding the Factors Associated with the Temporal Variability in Crash Severity before, during, and after the COVID-19 Shelter-in-Place Order », in : *Safety* vol. 8, no. 2, 42 p.

ADMINISTRATION, National Highway Traffic Safety (2021), « Continuation of Research on Traffic Safety During the COVID-19 Public Health Emergency : January–June 2021 », in : *NHTSA BSR Traffic Safety Facts*.

ADVANCEMENT OF AUTOMOTIVE MEDICINE, Association for the (s. d.), *Abbreviated Injury Scale*, https://www.aaam.org/abbreviated-injury-scale-ais-position-statement/, Récupéré le 06-03-2025.

AGRAWAL, N et DUHACHEK, A (2010), « Emotional Compatibility and the Effectiveness of Antidrinking Messages : A Defensive Processing Perspective on Shame and Guilt », in : *Journal of Marketing Research* vol. 47, no. 2, p. 263-273.

ANSELIN, L (1995), « Local Indicators of Spatial Association—LISA », in : *Geographical Analysis* vol. 27, no. 2, p. 93-115.

ASSUNCÃO, R et REIS, E (1999), « A new proposal to adjust Moran'sI for population density », in : *Statistics in Medicine* vol. 18, no. 16, p. 2147-2162.

BACHL, F (2024), *LGCPs - Multiple Likelihoods*, `https://inlabru-org.github.io/inlabru/articles/2d_lgcp_multilikelihood.html`, Récupéré le 30-04-2024.

BADDELEY, A, MØLLER, J et WAAGEPETERSEN, R (2001), « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns », in : *Statistica Neerlandica* vol. 54, no. 3, p. 329-350.

BADDELEY, A, RUBAK, E et TURNER, R (2015), *Spatial Point Patterns*, coll. Interdisciplinary Statistics, Philadelphia, USA : Chapman & Hall/CRC, 828 p.

BANERJEE, S, CARLIN, B, GELFAND, A et BANERJEE, S (2003), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman et Hall/CRC.

BENJAMINI, Y et HOCHBERG, Y (1995), « Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing », in : *Journal of the Royal Statistical Society*, coll. Series B (Methodological) vol. 57, no. 1, p. 289-300.

BENJAMINI, Y et YEKUTIELI, D (2001), « The control of the false discovery rate in multiple testing under dependency », in : *The Annals of Statistics* vol. 29, no. 4.

BERGMANN, T (2015), *Identifying outliers and influential cases*, `https://tillbe.github.io/outlier-influence-identification.html`, Récupéré le 30-12-2024.

BERNARDINELLI, L, CLAYTON, D, PASCUTTO, C, MONTOMOLI, C, GHISLANDI, M et SONGINI, M (1995), « Bayesian analysis of space—time variation in disease risk », in : *Statistics in Medicine* vol. 14, no. 21–22, p. 2433-2443.

BESAG, J (1974), « Spatial interaction and the statistical analysis of lattice systems », in : *Journal of the Royal Statistical Society* vol. 36, no. 2, p. 192-236.

BESAG, J, YORK, J et MOLLIÉ, A (1991), « Bayesian image restoration, with two applications in spatial statistics », in : *Annals of the Institute of Statistical Mathematics* vol. 43, no. 1, p. 1-20.

BIVAND, R, PEBESMA, E J et GOMEZ-RUBIO, V (2008a), *Applied spatial data analysis with R*, 2e éd., coll. Use R !, New York, USA : Springer.

— (2008b), *Applied spatial data analysis with R*, 1re éd., coll. Use R !, New York, USA : Springer.

BIVAND, R et WONG, D (2018), « Comparing implementations of global and local indicators of spatial association », in : *Test (Madrid)* vol. 27, no.3, p. 716-748.

BLANGIARDO, M et CAMELETTI, M (2015), *Spatial and Spatio-temporal Bayesian Models with R-INLA*, Wiley, 308p.

C, Marcia et SINGER, B (2006), « Controlling the False Discovery Rate : A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association », in : *Geographical Analysis* vol. 38, no. 2, p. 180-208.

CANADA, Association des industries de l'automobile du (2023), *Tendances des déplacements de véhicules : Les conducteurs canadiens maintiennent le cap au premier trimestre de 2023 | AIA Canada - Automotive Industries Association of Canada — aiacanada.com*, https://www.aiacanada.com/fr/nouvelles/tendances-des-deplacements-de-vehicules-les-conducteurs-canadiens-maintiennent-le-cap-au-premier-trimestre-de-2023/, Récupéré le 10-04-2024.

CANADA, Statistique (2022), *Ventes brutes d'essence au Canada, 2002 à 2022*, https://www150.statcan.gc.ca/n1/daily-quotidien/230919/cg-d001-fra.htm, Récupéré le 11-04-2024.

CANADA ET STREETLIGHT, Association des industries de l'automobile du (2022), *Quarterly Report : National Vehicle Kilometres Travelled Metrics - Q4 2021 and Q1 2022 (Partial)*, https://www.aiacanada.com/product/national-vehicle-kilometres-travelled-metrics-q4-2021-and-q1-2022-partial/, Récupéré le 10-04-2024.

CANCENSUS (s. d.), *Access, retrieve, and work with Canadian Census data and geography.* https://mountainmath.github.io/cancensus/, Récupéré le 07-03-2024.

CHAN, H, SKALI, A, SAVAGE, David A., STADELMANN, D et TORGLER, B (2020), « Risk attitudes and human mobility during the COVID-19 pandemic », in : *Scientific Reports* vol. 10, no. 1.

CMM (s. d.), *Données géoréférencées | Observatoire du Grand Montréal*, https://observatoire.cmm.qc.ca/produits/donnees-georeferencees/, Récupéré le 07-03-2024.

CRESSIE, N. (1993), *Statistics for spatial data*, 2e éd., coll. Probability & Mathematical Statistics S. Nashville, USA : John Wiley & Sons, 424 p.

Débit de circulation (s. d.), https://www.donneesquebec.ca/recherche/dataset/debit-de-circulation, Récupéré le 07-03-2024.

DIGGLE, P (2013), *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3e éd., Chapman et Hall/CRC.

DIGGLE, P, MORAGA, P, ROWLINGSON, B et TAYLOR, B (2013), « Spatial and Spatio-Temporal Log-Gaussian Cox Processes : Extending the Geostatistical Paradigm », in : *Statistical Science* vol. 28, no. 4, p. 542-563.

DIXON, P (2002), *Ripley's K function*, t. 3, John Wiley & Sons, p. 1796-1803.

DONG, X, XIE Xie, K et YANG, H (2022), « How did COVID-19 impact driving behaviors and crash Severity ? A multigroup structural equation modeling », in : *Accident Analysis and Prevention* vol. 172, p. 106687.

ELVIK, R, VAA, T, HOYE, A et SORENSEN, M (2009), *The Handbook of Road Safety Measures*, 2e éd., Emerald Group Publishing Limited, 1124p.

EMILIO, P, MORENO, B, ROBERT, S et O, Chris (2024), *The Matérn Model : A Journey through Statistics, Numerical Analysis and Machine Learning*.

ENTREPRISES, Laboratoire de données sur les (2024), *Workplace Mobility Tracker - Business Data Lab*, https://bdl-lde.ca/workplace-mobility-tracker/, Récupéré le 12-04-2024.

FOUNDATION, Traffic Injury Research (2023), *The Road Safety Monitor – Drinking and Driving - Traffic Injury Research Foundation*, https://tirf.ca/projects/road-safety-monitor-drinking-driving/, Récupéré le Accessed 13-04-2024.

FUGLSTAD, G, SIMPSON, D, LINDGREN, F et RUE, H (2019), « Constructing priors that penalize the complexity of Gaussian random fields », in : *Journal of the American Statistical Association* vol. 114, no. 525.

GÓMEZ-RUBIO, V (2020), *Bayesian Inference with INLA*, Chapman et Hall/CRC.

GONZALEZ, J et MORAGA, P (2023), « Non-parametric analysis of spatial and spatio-temporal point patterns », in : *The R journal* vol. 15, no.1, p. 65-82.

HUANG, H, ABDEL-ATY, M et DARWICHE, A (2010), « County-Level Crash Risk Analysis in Florida : Bayesian Spatial Modeling », in : *Transportation Research Record* vol. 2148, no. 1, p. 27-37.

INLA (s. d.), *inla.r-inla-download.org*, `https://inla.r-inla-download.org/r-inla.org/doc/latent/ar1.pdf`, Récupéré le 06-05-2024.

inla.r-inla-download.org (2020), `https://inla.r-inla-download.org/r-inla.org/doc/latent/seasonal.pdf`, Récupéré le 15-10-2024.

JOSH, P (2021), *Ottawa police see increase in impaired drivers on the roads in November*, `https://ottawa.ctvnews.ca/ottawa-police-see-increase-in-impaireddrivers-on-the-roads-in-november-1.5692850`, Récupéré le 08-04-2024.

KATRAKAZAS, C, MICHELARAKI, E, SEKADAKIS, M et YANNIS, G (2020), « A descriptive analysis of the effect of the COVID-19 pandemic on driving behavior and road safety », in : *Transportation Research Interdisciplinary Perspectives* vol. 7, p. 100186.

KNORR-HELD, L (2000), « Bayesian modelling of inseparable space-time variation in disease risk », in : *Statistics in medicine* vol. 19, no. 17-18, p. 2555-2567.

KRAINSKI, E et al. (2020), *Advanced spatial modeling with stochastic partial differential equations using R and INLA*, Philadelphia, USA : CRC Press.

LEROUX, B, LEI, X et BRESLOW, N (2000), *Estimation of Disease Rates in Small Areas : A new Mixed Model for Spatial Dependence*, sous la dir. d'E HALLORAN et D BERRY, Springer, p. 179-191.

LINDGRENN, F et BORCHERS, D (2024), *LGCPs - An example in two dimensions*, `https://inlabru-org.github.io/inlabru/articles/2d_lgcp_sf.html`, Récupéré le 29-04-2024.

LUM, C, MAUPIN, C et STOLTZ, M (2020), « The impact of COVID-19 on law enforcement agencies », in : *International Association of Chiefs of Police and George Mason University*.

LUM, Cynthia, MAUPIN, Carl et STOLTZ, Megan (2022), « The Supply and Demand Shifts in Policing at the Start of the Pandemic : A National Multi-Wave Survey of the Impacts of COVID-19 on American Law Enforcement », in : *Police Quarterly* vol. 26, no. 4, p. 495-519.

LYON, C, VANLAAR, W et ROBERTSON, R (2024), « The impact of COVID-19 on transportation-related and risky driving behaviors in Canada », in : *Transportation Research Part F : Traffic Psychology and Behaviour* 100, p. 13-21.

MAMRI, A et al. (2023), « Towards a comprehensive COVID-19 non-pharmaceutical interventions' index for the province of Quebec », in : *BMC research notes*, DOI : 10.21203/rs.3.rs-3500624/v1, URL : http://dx.doi.org/10.21203/rs.3.rs-3500624/v1.

MONTRÉAL, Service de police de la ville de (2023), *2022 - Rapport d'activités*, https://spvm.qc.ca/upload/Rapport_activites_2022_SPVM_Final.pdf, Récupéré le08-04-2024.

MORAGA, P (2023), *Spatial Statistics for Data Science*, coll. Data Science Series, Taylor & Francis, URL : https://www.paulamoraga.com/book-spatial/.

— (2019), *Geospatial Health Data : Modeling and Visualization with R-INLA and Shiny.* coll. Biostatistics Series, Chapman et Hall/CRC, URL : https://www.paulamoraga.com/book-geospatial/.

MORAN, P.A.P (1950), « Notes on Continuous Stochastic Phenomena », in : *Biometrika* vol 37, no. 1/2, p. 17-23.

NETO, T (2023), *QuebecTrafficVolumes*, https://github.com/nsaunier/OpenDataQuebec/tree/main/QuebecTrafficVolumes.

OPENSTREETMAP (s. d.), *OpenStreetMap*, https://www.openstreetmap.org, Récupéré le 07-03-2024.

PEBESMA, E et BIVAND, R (2023), *Spatial Data Science : With Applications in R*, Boca Raton, USA : CRC Press, 314p. URL : https://r-spatial.org/book/.

PHIL, H (2020), *Recent uptick in speeding on Edmonton roads concerns mayor and city official*, https://globalnews.ca/news/6739220/edmonton-drivers-speeding-coronavirus-covid-19-iveson/, Récupéré le 08-04-2024.

PROVINCE DE L'ONTARIO, comité d'examen des collisions mortelles (2020), *Conducteurs impliqués dans les collisions mortelles | Rapport annuel 2020 du Comité d'examen des collisions mortelles*, https://www.ontario.ca/fr/document/rapport-annuel-2020-du-comite-dexamen-des-collisions-mortelles-ottawa/conducteurs#~\protect\protect\leavevmode@ifvmode\kern+.2222em\relaxtext=Les%20comparaisons%20des%20conducteurs%20impliqu%C3%A9s,la%20population%20d'apr%C3%A8s%20les, Récupéré le 30-12-2024.

QUÉBEC, Province du (2024), *Rapports d'accident - Données Québec*, https://www.donneesquebec.ca/recherche/dataset/rapports-d-accident, Récupéré le 08-10-2024.

— (s. d.), *Découpages administratifs*, https://www.donneesquebec.ca/recherche/dataset/decoupages-administratifs, Récupéré le 07-03-2024.

QUÉBEC, Société de l'assurance automobile du (2021), *2020, bilan routier, faits saillants*, https://saaq.gouv.qc.ca/blob/saaq/documents/publications/bilan-routier-2020.pdf, Récupéré le 21-10-2024.

— (2022), *2021, bilan routier, faits saillants*, https://saaq.gouv.qc.ca/blob/saaq/documents/publications/bilan-routier-2021.pdf, Récupéré le 21-10-2024.

— (2023), *2022, bilan routier, faits saillants*, https://saaq.gouv.qc.ca/blob/saaq/documents/publications/bilan-routier-2022.pdf, Récupéré le 21-10-2024.

QUÉBEC, Sûreté du (2021), *Bilan routier de 20xx*, https://www.sq.gouv.qc.ca/communiques/devoilement-du-bilan-routier-2020/, Récupéré le 08-04-2024].

RASMUSSEN, C et WILLIAMS, C (2006), *Gaussian Processes for Machine Learning*, The MIT Press, 248 p.

RCMP (2022), *Rapport annuel de 2021*, https://www.rcmp-grc.gc.ca/nb/corporate-organisation/publications-manuals-publications-guides/2021-annual-report-rapport-annuel-de-2021-fra.htm, Récupéré le 08-04-2024.

— (2023), *Rapport annuel 2022*, https://www.rcmp-grc.gc.ca/nb/corporate-organisation/publications-manuals-publications-guides/2022-annual-report-rapport-annuel-de-2022-eng.htm, Récupéré le 08-04-2024.

SCHABENBERGER, O et GOTWAY, C (2005), *Texts in Statistical Science : Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, Taylor & Francis, 504 p.

SIMPSON, D, ILLIAN, J, LINDGREN, F, SORBYE, S et RUE, H (2016), « Going off grid : computationally efficient inference for log-Gaussian Cox processes », in : *Biometrika* vol. 103, no. 1, p. 49-70.

SPYCHALA, Cécile (2023), *Statistical analysis of road accidents in the region Franche-Comté : Risk factors for accident injuries and spatial modelling for accident occurrences*, URL : https://theses.hal.science/tel-04323459.

TELECOMMUNICATIONS, Cambridge Mobile (2021), *Measuring and Pricing Phone Distraction Risk*, https://www.cmtelematics.com/wp-content/uploads/2021/05/CMT_DDR_5-20_FINAL.pdf, Récupéré le 08-04-2024.

— (2023), *Cambridge Mobile Telematics reports increases in distracted driving caused an additional 420,000 crashes, 1,000 fatalities, and $10 billion in damages to the US economy in 2022 - Cambridge Mobile Telematics — cmtelematics.com*, https://www.cmtelematics.com/distracted-driving/cambridge-mobile-telematics-reports-increases-in-distracted-driving-caused-an-additional-420000-crashes-1000-fatalities-and-10-billion-in-damages-to-the-us-economy-in-2022/, Récupéré le 08-04-2024.

THIA, J (2021), *Fewer speeding tickets issued in 2020, but more drivers caught stunting and racing*, https://thestarphoenix.com/news/local-news/fewer-

speeding-tickets-issued-in-2020-but-more-drivers-caught-stunting-and-racing, Récupéré le 08-04-2024.

Tome V 2013 ; Signalisation routière - Volumes 1, 2 et 3 (2013), https://www.publicationsduquebec.gouv.qc.ca/produits-en-ligne/ouvrages-routiers/normes/collection-normes/tome-v-signalisation-routiere-volumes-1-2-et-3/, Récupéré le 09-12-2024.

TORONTO, Ville de (2020), *City of Toronto urges drivers to obey rules of the road*, https://www.toronto.ca/news/city-of-toronto-urges-drivers-to-obey-rules-of-the-road/, Récupéré le 08-04-2024.

TRANSPORTATION, US Department of et ADMINISTRATION, Federal Highway (2018), *Traffic Data, Computation Method, Pocket guide*, https://www.fhwa.dot.gov/policyinformation/pubs/pl18027_traffic_data_pocket_guide.pdf, Récupéré le 27-11-2024.

TRANSPORTATION STATISTICS, Bureau of (2022), *Roadway Vehicle-Miles Traveled (VMT) and VMT per Lane-Mile by Functional Class*, https://www.bts.gov/content/roadway-vehicle-miles-traveled-vmt-and-vmt-lane-mile-functional-class, Récupéré le 12-04-2024.

VANLAAR, W et al. (2021), « The impact of COVID-19 on road safety in Canada and the United States », in : *Accident Analysis & Prevention* 160, p. 106324.

WHITTLE, P (1963), « Stochastic processes in several dimensions », in : *Bulletin of the International Statistical Institute* vol. 40, p. 974-994.

WIKLE, C, ZAMMIT-MANGION, A et CRESSIE, N (2019), *Spatio-temporal statistics with R*, coll. The R series, Boca Raton, USA : CRC Press.

Annexe A – Manquement éthique du Comité d'Éthique de la Recherche (CER) du HEC

Nous avons reçu un manquement éthique concernant l'utilisation des données dans le cadre de ce projet. Toutefois, il est important de préciser que ce manquement était hors de notre contrôle, car nous croyions être en conformité avec les règles et agissions de bonne foi tout au long de l'analyse. Nous avions reçu l'autorisation de travailler avec ces données et, à l'époque, les procédures administratives semblaient respecter les exigences éthiques en vigueur. De plus, une entente avait déjà été signée directement avec la SAAQ. Ce n'est que plus tard qu'il a été constaté qu'une certification éthique spécifique était nécessaire, bien que nous n'en ayons pas été informés initialement. Cependant, comme l'indique l'annexe 2 de ce mémoire, nous avons finalement obtenu le droit d'utiliser les données et de publier ce travail, ce qui montre que toutes les démarches ont été régularisées a posteriori.

Comité d'éthique
de la recherche

HEC MONTRÉAL

Montréal, le lundi 2 décembre 2024

Monsieur Edgar Lanoue
Étudiant,
M. Sc. en science des données et analytique d'affaires
HEC Montréal

Madame Aurélie Labbe
Professeure titulaire,
Département de sciences de la décision
HEC Montréal

Objet : Manquement à la Politique relative à l'éthique de la recherche avec des êtres humains - Projet intitulé « COVID-19, comportements et sécurité routière : Analyses spatiales et temporelles des accidents de la route au Québec »

Monsieur Edgar Lanoue,
Madame Aurélie Labbe,

Le CER de HEC Montréal a récemment pris connaissance du fait que la collecte de données, réalisée dans le cadre du projet mentionné en objet, a été menée avant d'avoir obtenu l'approbation formelle du CER.

Comme vous le savez, le CER de HEC Montréal a le mandat de s'assurer que les projets qui y sont menés respectent la politique de l'École sur la conduite de la recherche auprès de sujets humains et les principes établis dans l'Énoncé de politique des trois conseils : *Éthique de la recherche avec des êtres humains* (EPTC2), politique fédérale en éthique de la recherche. Comme spécifié à l'Article 2.1 de l'EPTC2, les recherches avec des participants humains doivent faire l'objet d'une évaluation de l'éthique et être approuvées par un CER **avant** le début de la collecte de données.

Puisque l'Article 2.1 de l'EPTC2 n'a pas été respecté, le comité d'éthique de la recherche de HEC Montréal ne peut émettre rétroactivement la certification éthique demandée, pas plus qu'il ne peut délivrer d'*Attestation d'approbation éthique complétée*.

Cette attestation étant nécessaire à l'obtention du diplôme de maîtrise, nous vous invitons à communiquer avec Mme Sihem Taboubi, directrice des programmes de M. Sc. à HEC Montréal, pour prendre connaissance des procédures à suivre.

Nous souhaitons également vous rappeler qu'il est de la responsabilité conjointe de l'étudiant ou de l'étudiante et de son directeur ou de sa directrice de s'assurer d'obtenir une approbation éthique formelle auprès du CER de HEC Montréal avant d'entamer toute collecte ou consultation de données impliquant des personnes.

Veuillez recevoir, Monsieur, Madame, mes salutations distinguées.

Le président du Comité d'éthique de la recherche de HEC Montréal

Maurice Lemelin

c. c. Sihem Taboubi
Michèle Breton
Caroline Aubé

Annexe B – Lettre du Comité d'Éthique de la Recherche (CER) de Polytechnique, autorisant l'utilisation des données dans le cadre de la recherche

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Montréal, le 13 décembre 2024

PAR COURRIEL

M. Nicolas Saunier
Professeur titulaire
Département des génies civil, géologique et des mines

Objet: Utilisation des données et des renseignements personnels en l'absence d'une approbation éthique – CER-2324-64-M

Monsieur Saunier,

La présente fait suite à la situation de manquement relatif à votre projet intitulé « COVID-19, comportements et sécurité routière: bilan et leçons d'un bouleversement » et financé par le FRQ (dossier 322056). Comme vous le savez, la personne chargée de la conduite responsable en recherche (PCCRR) de Polytechnique Montréal a confirmé qu'il y a eu manquement en matière d'éthique de la recherche avec des êtres humains, notamment en regard des articles 5.5a (utilisation secondaire de renseignements identificatoires sans le consentement des personnes) et 6.11 (obtention d'une approbation éthique avant la consultation des données) de l'Énoncé de politique des trois Conseils (2022). Ce manquement a été dûment signalé au FRQ et cet organisme subventionnaire a cautionné nos recommandations.

Dans la mesure où le manquement est qualifié de mineur par la PCCRR et qu'aucun participant n'a directement subi de préjudice lié à la situation, le Comité d'éthique de la recherche de Polytechnique Montréal vous a informé en juin 2024 qu'elle levait l'interdiction d'utilisation des données à des fins de recherche et vous permettait de les utiliser selon les conditions prévues à l'entente de transfert ratifiée avec la SAAQ et sanctionnée par la CAI. Il importe de souligner que cette autorisation d'utilisation ne constitue pas une approbation éthique en soi puisque le CER ne peut émettre a posteriori une telle approbation (cf. EPTC, 2022 : art. 6.11). De manière concomitante, nous ne nous opposons pas à la publication des résultats découlant du projet de recherche.

Dans le cadre de ce projet, vous assumiez la codirection d'un étudiant inscrit à HEC Montréal, M. Edgar Lanoue. Bien que cet étudiant ne soit pas sous notre autorité, mais qu'il réalise de la recherche sous nos auspices et avec les données visées par le manquement, les conditions d'utilisation des données s'étendent à ses travaux si tant est que leur utilisation est permise par l'entente qui vous lie au fournisseur de données et selon les conditions énoncées par ce dernier.

En espérant que ces informations vous éclairent, nous vous prions, Monsieur Saunier, d'agréer l'expression de nos plus cordiales salutations.

Philippe Doyon-Poulin

Philippe Doyon-Poulin
Président du comité d'éthique de la recherche
Professeur agrégé
Département de mathématiques et de génie industriel

Guillaume Paré
Directeur et PCCRR
Bureau de l'éthique et de l'intégrité en recherche (BEIR)

Comité d'éthique de la recherche
Bureau de l'éthique et de l'intégrité en recherche (BEIR)
Pavillon principal
Téléphone : (514) 340-4711, poste 3830
Courriel : ethique@polymtl.ca

Adresse postale
C.P. 6079, succ. Centre-ville
Montréal (Québec) Canada H3C 3A7

Adresse physique
2500, chemin de Polytechnique
Montréal (Québec) Canada H3T 1J4