UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

MÉTHODES BASÉES SUR L'IA ET DONNÉES LIDAR POUR L'ÉLABORATION D'INDICATEURS DE PERFORMANCE DE LA FLUIDITÉ ET DE LA SÉCURITÉ ROUTIÈRE

MÉMOIRE PRÉSENTÉ

COMME EXIGENCE PARTIELLE DE LA MAÎTRISE EN GÉNIE ÉLECTRIQUE

PAR

MAME THIERNO MBACKÉ FALL

DÉCEMBRE 2025

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

MAÎTRISE EN GÉNIE ÉLECTRIQUE (M. Sc. A.)

Direction de recherche :

Daniel Massicotte	Directeur de recherche
Messaoud Ahmed Ouameur	Co-Directeur de recherche

Jury d'évaluation

Fadel TOURÉ, UQTR	Membre interne
Daniel Massicotte, UQTR	Directeur de recherche
Mohammed Bahoura,UQAR	Évaluateur externe

Abstract

Safety and fluidity are major concerns in modern road traffic infrastructures. This thesis focuses on these key aspects of road traffic engineering by proposing methods for improved traffic monitoring and risk evaluation at intersections.

Monitoring traffic intersections is essential for anticipating peak periods and assessing safety risks in order to take timely action to optimize flow and reduce hazardous situations. With recent advances in sensor technologies, particularly Lidar-based systems, it is now possible to obtain precise vehicle counts and classify vehicle types effectively. In this work, traffic user counts and Post-Encroachment Time (PET) are used as key surrogate safety measures. These indicators respectively quantify traffic volume and the time interval between two road users crossing the same conflict point, thereby providing insight into potential collision risks.

The first part of this thesis proposes a short-term traffic flow prediction method using count data. Three models are compared: SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, a hybrid SARIMA model, and an Adaptive Extended Kalman Filter (AEKF) with Recursive Least Squares (RLS). The results demonstrate that the hybrid SARIMA model achieves the best performance in terms of RMSE (Root Mean Squarred Error), MAE (Mean Absolute Error), and R^2 coefficient.

In the second part, PET data are used to develop a risk index based on a hybrid method. First, the DBSCAN (Density Based Spatial Clustering Applications with Noises) algorithm identifies critical anomalies (extreme PET values), which are then used to define bivariate thresholds. These thresholds guide the fitting of a statistical distribution. Next, several supervised learning models—including optimized SVM and neural networks—are evaluated for risk classification.

This twofold approach contributes to enhancing both traffic prediction and proactive risk assessment at intersections, supporting the development of smarter and safer road infrastructures.

Keywords : Surrogate measures, road traffic safety, clustering, risk index, classification, flux prediction

Résumé

La sécurité et la fluidité constituent des enjeux majeurs dans les infrastructures routières modernes. Ce mémoire porte sur ces aspects essentiels de l'ingénierie du trafic routier, en proposant des méthodes d'analyse pour une meilleure surveillance et évaluation du risque aux intersections.

La surveillance des intersections est cruciale pour anticiper les heures de pointe, évaluer les niveaux de risque, et agir de manière à améliorer la fluidité tout en réduisant les situations dangereuses. Grâce aux récents progrès des technologies de capteurs, notamment les systèmes basés sur le LiDAR, il est désormais possible d'obtenir un comptage précis des véhicules et de les classifier efficacement. Ce travail s'appuie sur deux mesures substitutives clés : le comptage des usagers de la route et le temps post-enchevêtrement (PET). Le comptage renseigne sur le volume de trafic, tandis que le PET mesure le temps entre le passage de deux usagers sur un même point de conflit, donnant ainsi une indication du risque de collision.

Dans un premier temps, ce mémoire présente une méthode de prévision du flux de véhicules à court terme à partir des données de comptage. Trois modèles sont comparés : le modèle SARIMA, le modèle SARIMA hybride, et le filtre de Kalman étendu adaptatif (AEKF) associé à la méthode des moindres carrés récursifs (RLS). Les résultats expérimentaux montrent que le modèle hybride SARIMA est le plus performant selon les indicateurs RMSE, MAE et le coefficient de détermination R^2.

Dans un second temps, les données PET sont exploitées pour développer un indice de risque à l'aide d'une approche hybride. L'algorithme de regroupement DBSCAN est utilisé pour détecter les valeurs extrêmes (anomalies), à partir desquelles des seuils bivariés sont définis afin d'ajuster une fonction de distribution. Ensuite, plusieurs modèles

d'apprentissage supervisé, tels que les SVM (*Support Vector Machine*) optimisés et les réseaux de neurones, sont évalués pour la classification du risque.

Cette approche en deux volets contribue à améliorer à la fois la prévision du trafic et l'évaluation proactive des risques aux intersections, soutenant ainsi le développement d'infrastructures routières plus intelligentes et sécurisées.

Mots clés : mesures de substitution, sécurité routière, regroupement, classification, indice de sécurité, prédiction de flux.

Avant-propos

Ce mémoire représente l'aboutissement d'un parcours académique et humain riche, marqué par de nombreux défis et de précieuses rencontres. Il est le fruit d'une volonté profonde de contribuer à l'amélioration des conditions de vie au sein de nos communautés, notamment par le biais de technologies innovantes et durables dans le domaine des transports. Ce projet s'inscrit dans la dynamique de transformation vers des infrastructures plus sûres, plus fluides et plus intelligentes.

Je tiens à exprimer ma profonde gratitude envers toutes celles et tous ceux qui ont rendu ce travail possible. Mes remerciements chaleureux vont aux professeurs Daniel Massicotte du Laboratoire de signaux et systèmes intégrés, Messaoud Ahmed Ouameur du département de génie électrique, ainsi qu'à Jean-Sébastien Dessureault du département de mathématiques et informatique de l'Université du Québec à Trois-Rivières, pour leurs conseils éclairés et leurs orientations précieuses.

Un immense merci à la Ville de Trois-Rivières, et en particulier à Vincent Turgeon, dont le soutien indéfectible a permis à ce projet de se concrétiser. À l'équipe technique de Velodyne Blue City, pour leur apport technique et leur accompagnement dans l'exploitation de la technologie LiDAR. Enfin, mes pensées les plus sincères vont à Cheikh AB et Coumba, mes parents bien-aimés, à mon ex-conjointe, à mes deux merveilleux enfants, et à la mémoire de mon frère défunt Mor Fall, ainsi que des regrettés Oumar Thiam et Mouhamadou Moustapha Sall, à qui je dédie humblement ce mémoire.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ACF	Auto Correlation Function
AEKF	Adaptive Extended Kalman Filter
AR	Auto Regressive
ARIMA	Auto Regressive Integrated Moving Average
DBSCAN	Density Based Spatial Clustering of Applications with Noise
EBLT	East Bound Left Turn
EBRT	East Bound Right Turn
EBST	East Bound Straight Through
EKF	Extended Kalman Filter
elr	East Left to Right
ELR	Efficient Linear Regression
erl	East Right to Left
FNR	False Negative Rate
GNB	Gaussian Naive Bayes
GPD	Generalized Pareto Distribution
LiDAR	Light Detection and Ranging
LMS	Least Mean Squares
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
NBLT	North Bound Left Turn
NBRT	North Bound Right Turn
NBST	North Bound Straight Through

nlr	North Left to Right
nrl	North Right to Left
ONN	Optimized Neural Network
OSVM	Optimized Support Vector Machine
PACF	Partial Auto Correlation Function
PET	Post-Encroachment Time
ReLU	Rectified Linear Unit
RLS	Root Least Square
RMSE	Root Mean Squared Error
SARIMA	Seasonal Auto Regressive Integrated Moving Average
SBLT	South Bound Left Turn
SBRT	South Bound Right Turn
SBST	South Bound Straight Through
slr	South Left to Right
srl	South Right to Left
tanh	Tangente hyperbolique
TPA	Temps Post-Accrochage
TPR	True Positive Rate
TTC	Time to Collision
WBLT	West Bound Left Turn
WBRT	West Bound Right Turn
WBST	West Bound Straight Through

wlr	West Left to Right
wrl	West Right to Left

Chapitre 1 - Introduction

1.1 Contexte général

Chaque jour, des milliards d'individus empruntent les routes pour se rendre à leurs activités quotidiennes. Des infrastructures routières de qualité participent grandement à la vie active des populations, améliorent leurs conditions de vie et représentent un important levier économique. La qualité de ces infrastructures dépend de plusieurs facteurs, dont les trois principaux sont la sécurité, la mobilité et la fluidité ainsi que l'environnement et l'écologie.

Chaque année, 1,35 million de personnes sont tuées sur les routes à l'échelle mondiale, et entre 20 et 50 millions de personnes sont blessées sérieusement [3]. Les projections pour 2030 indiquent que le nombre de victime (blessés et tués) pourrait atteindre 265 millions de personnes en l'absence de solutions efficaces.

Selon une étude de Statistique Canada [4], la congestion du réseau routier, en particulier dans les régions urbaines, entraîne des répercussions importantes sur tous les usagers du système de transport, avec des impacts économiques, environnementaux et sanitaires significatifs.

Le manque de fluidité et de mobilité urbaine augmente les risques de pollution atmosphérique. Les émissions de gaz à effet de serre, notamment le CO_2, sont destructrices pour la couche d'ozone, contribuant ainsi au réchauffement climatique et aux dérèglements environnementaux. Face à ces enjeux, les municipalités sont confrontées à d'importants défis et comptent sur les technologies existantes et émergentes pour y répondre.

Ce projet s'inscrit dans le cadre d'une collaboration avec la Ville de Trois-Rivières. Son objectif est de développer une méthode, basée sur l'intelligence artificielle (IA), pour

l'élaboration d'indicateurs pertinents dans les intersections urbaines majeures de la ville. Grâce à des capteurs *LiDAR*, des données sont collectées sur la circulation des piétons, vélos, voitures, camions et autres usagers. Cette masse de données représente une opportunité pour concevoir des outils de mesure permettant d'évaluer l'état des intersections critiques de la ville sous l'angle de la sécurité, de la mobilité et de l'environnement.

Le projet constitue une avancée importante pour la Ville de Trois-Rivières, facilitant sa transition vers une ville plus intelligente, durable et sécuritaire. Il s'intègre dans les travaux de la Chaire de recherche sur les signaux et l'intelligence des systèmes haute performance ainsi que dans ceux du Laboratoire des signaux et systèmes intégrés.

Ce travail s'inscrit dans le cadre de la mise en œuvre de la politique de Tolérance Zéro, une stratégie municipale qui vise à éliminer les accidents graves et mortels sur le réseau routier par une approche intégrée de prévention, de détection des comportements à risque, et d'intervention rapide. Cette politique repose sur l'exploitation de données massives pour améliorer la sécurité urbaine et anticiper les situations à haut risque.

1.2 Problématique

Parmi les solutions techniques actuellement disponibles pour les municipalités, aucune ne permet de connaître, de manière adéquate et en temps réel, le nombre de véhicules en déplacement, de bus ou de piétons présents sur une courte période. Les boucles de détection installées aux intersections permettent certes d'identifier des pics ponctuels de volume d'activités de mobilité urbaine, mais elles présentent plusieurs limitations, notamment en ce qui concerne la capacité à fournir des informations sur les interactions entre véhicules.

Dans une volonté d'innovation, la Ville de Trois-Rivières a mis en place un réseau de capteurs *LiDAR* aux intersections stratégiques. Ce réseau vise à alimenter un système d'IA

permettant de prédire les patrons de sécurité et de volume, ainsi que d'identifier les zones à risque élevé.

1.3 Objectifs

L'objectif principal du présent projet est de développer une méthode à base d'apprentissage automatique, un sous-domaine de l'IA, permettant de générer des indicateurs pertinents pour évaluer, en temps quasi réel, la sécurité et la fluidité des intersections urbaines à partir de données issues de capteurs LiDAR.

Cet objectif s'articule autour des sous-objectifs suivants :

- améliorer l'efficacité du déploiement des ressources municipales afin de rendre les routes plus sécuritaires, plus fluides et plus écologiques;
- implémenter et tester des techniques émergentes, telles que l'IA, dans les domaines du transport et de l'ingénierie routière
- identifier, à partir de l'observation quotidienne de l'activité urbaine, des patrons et des marqueurs pertinents en matière de sécurité et de fluidité
- prédire, sur des horizons temporels rapprochés, le niveau de risque ou le volume d'usagers

La mobilité urbaine commence à offrir d'intéressantes perspectives. Nos travaux ciblent les aspects autant dans la fluidité que dans la sécurité et présentent une particularité de fournir, en temps réel : des prédictions à court terme pour la fluidité en appliquant des modèles conventionnels comme la méthode de Kalman étendue ainsi que l'identification des heures et des jours à haut risque pour la sécurité des usagers dans un souci de prévention afin de mieux orienter les services de contrôle de la sécurité de la municipalité.

1.4 Méthodologie

Afin d'atteindre les objectifs du projet, une méthodologie structurée en trois grandes étapes a été adoptée. Elle repose sur une combinaison de recherche bibliographique, de développement algorithmique, et de validation expérimentale en contexte réel.

Revue de littérature et analyse technologique : consiste en une recherche bibliographique approfondie portant sur les technologies de collecte et d'analyse de données liées au trafic routier, en particulier celles basées sur les capteurs *LiDAR*. Cette revue permet de : cerner les capacités techniques des capteurs LiDAR en matière de détection, de comptage et de classification d'usagers de la route (piétons, cyclistes, véhicules motorisés, etc.); identifier les limites technologiques et pratiques de ces dispositifs (sensibilité aux conditions météorologiques, coût, résolution, fréquence d'échantillonnage, etc.); répertorier les modèles d'exploitation et les indicateurs déjà utilisés dans la littérature pour évaluer la performance du trafic et la sécurité routière. Des exemples concrets issus de publications scientifiques (notamment [5]) illustrent les usages actuels du LiDAR dans le domaine du transport intelligent.

Développement d'une méthode d'analyse basée sur l'apprentissage automatique : concerne l'élaboration d'un cadre analytique intégrant des techniques d'apprentissage automatique appliquées aux données collectées. Trois principales taches sont réalisées par l'apprentissage automatique : le clustering, la classification et la régression.

- **Méthodes non supervisées** : essentiellement basées sur le clustering (comme DBSCAN), elles servent à mettre en évidence des regroupements naturels dans les indicateurs de sécurité — par exemple des valeurs atypiques du Temps Post-Accrochage — ou à identifier des anomalies ;

- **Méthodes supervisées de classification** : comprenant notamment les réseaux de neurones, les SVM et les modèles non linéaires à mémoire (NARX), elles permettent d'estimer le niveau de risque à partir de caractéristiques extraites des données brutes (vitesse, trajectoire, type d'usager, etc.) ;
- **Méthodes fondées sur les séries temporelles** : telles que les modèles SARIMA (*Seasonal AutoRegressive Integrated Moving Average*), les filtres de Kalman adaptatifs et les approches RLS (*Recursive Least Squares*), elles sont utilisées pour la prévision à court terme des volumes d'usagers en fonction des heures, des jours ou de périodes particulières.

Ces approches sont sélectionnées en fonction de la nature des jeux de données (temporelles ou événementielles) et des indicateurs ciblés (sécurité, fluidité).

Validation expérimentale en contexte de données LiDAR réelles: est dédiée à la validation des méthodes développées, à partir de données massives recueillies en contexte réel dans une quinzaine d'intersections à Trois-Rivières, équipées de capteurs *LiDAR*. Cette phase vise à : a) évaluer la robustesse des algorithmes face aux données réelles (variabilité, bruit, données manquantes); b) comparer les performances des différentes approches en termes de précision, de capacité prédictive et de temps de traitement; c) proposer des indicateurs synthétiques exploitables par les gestionnaires municipaux pour orienter les interventions (sécurisation, fluidification, planification des horaires, etc.).

1.5 Contributions

Notre contribution est double. D'une part, elle renforce la proactivité des services municipaux en leur fournissant des outils d'analyse prédictive fondés sur l'apprentissage automatique, capables d'identifier en temps réel les zones ou les périodes à risque accru.

D'autre part, elle améliore la capacité des autorités à planifier des interventions ciblées et à adapter l'infrastructure routière de manière dynamique en fonction des comportements observés et des tendances détectées.

Sur le plan scientifique, cette recherche représente une avancée significative dans le domaine de la modélisation intelligente du trafic urbain. Elle introduit une méthode innovante d'exploitation de données *LiDAR* issues d'intersections urbaines, en combinant l'apprentissage automatique, la théorie des valeurs extrêmes et les techniques de détection d'anomalies. L'approche proposée permet non seulement de quantifier le niveau de risque routier avec une grande précision, mais aussi de proposer des indicateurs robustes et interprétables pour la prise de décision. Ces avancées majeures sont sanctionnées par leur présentation à la conférence IEEE *Digital Signal Processing 2025* à Costa Navarino en Grèce en juin 2025 [6].

En somme, ce travail constitue une valeur ajoutée stratégique tant pour la recherche appliquée en transport intelligent que pour les municipalités engagées dans une transition vers des villes intelligentes, sûres et durables et répond aux défis de la politique de Tolérance Zéro.

1.6 Organisation du mémoire

Après ce chapitre introductif, le chapitre 2 présente une revue de la littérature traitant des études scientifiques appliquées au transport intelligent. Le chapitre 3 propose ensuite une analyse des données disponibles. Le chapitre 4 décrit en détail la méthodologie retenue. Enfin, le chapitre 5 discute des résultats obtenus.

Chapitre 2 - Revue de la littérature

Ce chapitre présente une revue de la littérature pertinente au projet, afin de situer les fondements théoriques et technologiques sur lesquels repose l'approche développée. Il a pour objectif de contextualiser les travaux en mettant en lumière les solutions existantes, les approches méthodologiques et les résultats obtenus par la communauté scientifique en lien avec la sécurité et la fluidité du trafic routier.

Dans un premier temps, nous décrivons les technologies de collecte de données de trafic, en particulier les capteurs *LiDAR*, qui constituent la base de notre acquisition de données. Ensuite, nous présentons la notion de performance du trafic routier et les indicateurs généralement utilisés pour l'évaluer, en insistant sur les dimensions de sécurité et de fluidité. Enfin, nous analysons les travaux scientifiques antérieurs portant sur l'utilisation de données de trafic, notamment les approches fondées sur les mesures indirectes de sécurité et les apports récents de l'IA dans ce domaine. Cette analyse permettra de mettre en évidence les lacunes existantes dans la littérature et de justifier la pertinence et l'originalité de la démarche proposée dans le cadre de cette étude.

2.1 Technologie *LiDAR* pour la collecte de données de trafic

De nombreuses technologies ont été expérimentées au fil des années pour répondre aux enjeux de sécurité et de fluidité du trafic routier. Des boucles de détection inductives aux caméras de surveillance, en passant par les radars à ondes millimétriques, une nouvelle génération de capteurs, notamment le *LiDAR*, est aujourd'hui en plein essor.

Dans le cadre de cette étude, les données de trafic ont été fournies par un prestataire externe via des capteurs *LiDAR Velodyne*, opérés par *Blue City*. Nous ne nous attardons

pas ici sur les détails techniques d'acquisition et de traitement préliminaire des données, celui-ci étant entièrement pris en charge par l'opérateur.

Le *LiDAR* présente plusieurs avantages notables qui peuvent compenser son principal inconvénient : son coût relativement élevé. En générant des nuages de points à haute résolution, cette technologie permet l'extraction d'informations fines sur l'environnement : contours et formes des objets, état des chaussées, formes architecturales, etc. Il est ainsi largement utilisé pour la modélisation 2D et 3D de scènes urbaines complexes [7].

Les données issues du *LiDAR* peuvent être exploitées pour l'aménagement des corridors de circulation, l'analyse du flux de trafic, l'évaluation de la sécurité routière (nivellement, dévers), la protection des infrastructures critiques ou bien le repérage des câbles, obstacles et éléments urbains.

Cependant, comme toute technologie optique, le *LiDAR* est sensible aux conditions météorologiques. Il est notamment inefficace en présence de brouillard dense ou de nuages épais [7], ce qui peut limiter sa capacité de collecte de données dans certaines situations climatiques.

L'étude [8] illustre une application concrète du *LiDAR* sur un tronçon routier dans le Missouri. Elle évalue l'efficacité de divers logiciels de traitement et d'extraction d'information à partir des nuages de points *LiDAR*, en mettant en évidence à la fois les potentiels (notamment du *LiDAR* mobile) pour améliorer la collecte de données et la réduction des risques, mais aussi les limites actuelles liées aux outils de traitement. Le *LiDAR* est une technologie non intrusive de la vie privée contrairement aux caméras CCTV.

2.2 Notions de performance d'un trafic routier

La performance du trafic routier désigne la capacité d'un réseau à assurer un déplacement efficace des véhicules et des usagers en minimisant les temps de parcours et en réduisant les risques d'accidents et les impacts environnementaux. Une circulation fluide et bien régulée est essentielle pour améliorer la mobilité urbaine et limiter les coûts économiques et sociaux liés aux embouteillages.

La performance du trafic routier est un enjeu majeur dans l'aménagement urbain et la gestion des infrastructures de transport. Elle permet d'évaluer l'efficacité d'un réseau routier en termes de fluidité, de sécurité et d'impact environnemental. Cette section propose une analyse des principaux indicateurs de performance du trafic routier, ainsi que des facteurs influençant cette performance. Contraints et limités par nos données, nous faisons par la suite une synthèse en fonction des ressources disponibles dans notre cas d'exemple, car tout ne peut être appliqué et ne fait pas l'objet de nos travaux.

L'évaluation de la performance du trafic repose sur plusieurs indicateurs quantitatifs et qualitatifs :

Débit est le nombre de véhicules passant en un point donné d'une route durant un intervalle de temps, généralement exprimé en véhicules par heure. Il permet de mesurer la capacité d'une voie de circulation.

Vitesse moyenne est un indicateur clé qui reflète la fluidité du trafic. Elle est influencée par la densité du trafic, la réglementation et les conditions de la route.

Densité du trafic est définie comme le nombre de véhicules présents sur une unité de longueur de la route (véhicules/km). Une densité élevée peut indiquer une congestion.

Temps de parcours moyen sur un tronçon donné est un facteur essentiel dans l'analyse de la performance du trafic, notamment pour les usagers réguliers.

Taux d'occupation est un indicateur qui représente le pourcentage de la capacité d'une route utilisée à un instant donné.

Sécurité routière: Le taux d'accidents par unité de trafic permet d'évaluer la sûreté d'un réseau routier. Une augmentation du nombre d'accidents peut indiquer des problèmes structurels ou de gestion.

Émissions polluantes : Les embouteillages entraînent une augmentation des émissions de gaz à effet de serre (CO_2, NOx, particules fines), impactant la qualité de l'air.

2.3 Travaux scientifiques

Dans le domaine des systèmes de transport, de nombreux travaux ont déjà été réalisés, mais peu d'entre eux abordent le problème crucial de la prévention de la sécurité de manière efficace pour soutenir une ingénierie de gestion du trafic appropriée. L'utilisation de l'IA dans les systèmes de transport devient de plus en plus importante. Les systèmes de transport intelligents se développent et intègrent des méthodes d'analyse et de traitement des données pour répondre à la forte demande de préoccupations économiques, de fiabilité et de qualité des infrastructures et surtout de sécurité des usagers de la route.

En effet, des milliards de personnes utilisent chaque jour les infrastructures routières pour se rendre sur les lieux de leurs activités quotidiennes. Une infrastructure routière de qualité joue un rôle important dans la vie professionnelle des personnes, améliore leurs conditions de vie et représente un levier économique important. La qualité des infrastructures dépend

de plusieurs facteurs, et la question de la sécurité reste l'une des plus importantes, voire cruciale.

Du point de vue de la fluidité, plusieurs études ont été menées en ce qui a trait à la prédiction de flux. [9] propose une nouvelle méthode hybride d'intelligence artificielle pour améliorer la prévision à court terme du trafic urbain dans les réseaux artériels avec feux tricolores. Une variante du filtre de Kalman, appelée *Noise-Immune Kalman Filter* (NIKF), afin d'améliorer la prévision à court terme du trafic routier, même en présence de bruits non-gaussiens, fréquents dans les données réelles, est proposée par [10]. Les auteurs de [11] ont développé une méthode de prévision du trafic à court terme capable de prédire le niveau de trafic (nombre de véhicules par intervalle de temps), de quantifier l'incertitude associée à cette prévision (intervalle de confiance). Deux modèles de prévision dynamique du trafic à court terme en utilisant la théorie du filtre de Kalman, afin d'améliorer la précision des estimations du volume de trafic sur les réseaux urbains, ont été développés par [12].

Au niveau de l'aspect sécuritaire, les auteurs dans [13] nous apprennent que les mesures indirectes de sécurité offrent une alternative précieuse lorsque les données d'accidents sont indisponibles, insuffisantes ou inadaptées, notamment dans le cas d'infrastructures récentes ou atypiques. Leur principal avantage réside dans la possibilité d'agir préventivement, sans attendre la survenance d'un nombre significatif d'accidents. En analysant les événements précurseurs des collisions, ces mesures permettent de détecter plus rapidement les zones à risque et de proposer des interventions ciblées.

Une approche d'analyse de sécurité basée sur les accidents se heurte également à plusieurs lacunes, telles que le caractère aléatoire et la rareté des accidents, le manque de réactivité et l'incohérence des rapports d'accidents [1].

Cela dit, les mesures de sécurité indirectes peuvent jouer un rôle clé dans la prévention des accidents car si les facteurs qui ont contribué aux accidents sont identifiés, il est alors possible de modifier et d'améliorer le système de transport [2], ce qui rend le niveau de sécurité du trafic plus intéressant à traiter.

De nombreuses études antérieures ont proposé plusieurs méthodes pour évaluer la sécurité en fonction des activités de trafic au lieu d'utiliser des données historiques sur les accidents et les collisions qui nécessitent une longue période de collecte de données. Nous notons ensuite une augmentation des mesures de substitution pour assurer l'efficacité de la prévention par une intervention en temps réel et une gestion précise des ressources de contrôle de la sécurité. [14] propose une méthode d'évaluation de la sécurité routière basée sur les conflits en associant la fréquence et la gravité des conflits aux caractéristiques du trafic à court terme. [15] effectue une revue systématique des mesures de sécurité basées sur les conflits en mettant l'accent sur le contexte de leurs applications. [16] utilise des indicateurs de conflit, le TPA (temps post-accrochage) et TTC (*Time To Collision*), qui sont expliqués au chapitre 3, pour identifier les conflits entre piétons et le modèle proposé peut prédire les conflits entre piétons un cycle à l'avance, ce qui peut être de 2 à 3 minutes. [17] a utilisé des indicateurs de sécurité de substitution pour mesurer le niveau de sécurité des conflits entre piétons et autres usagers de la route afin d'évaluer le risque de conflit.

2.4 Conclusion

Ce chapitre a permis de dresser un panorama complet des fondements technologiques et méthodologiques en lien avec la sécurité et la fluidité du trafic routier. Trois axes majeurs ont été explorés pour contextualiser la démarche adoptée dans cette étude :

1. Les technologies de collecte des données, avec une attention particulière portée sur le *LiDAR*, ont démontré leur pertinence pour la modélisation fine de l'environnement urbain. Malgré certaines limites, notamment liées aux conditions météorologiques, le *LiDAR* s'impose comme un outil non intrusif et efficace pour enrichir les analyses de trafic [18].
2. La notion de performance du trafic routier, abordée à travers des indicateurs tels que le débit, la vitesse moyenne, la densité, le temps de parcours, le taux d'occupation, la sécurité routière et les émissions polluantes. Ces indicateurs constituent des leviers essentiels pour diagnostiquer et améliorer la qualité des réseaux routiers.
3. Les apports de la recherche scientifique, en particulier l'usage croissant de l'intelligence artificielle et des mesures indirectes de sécurité. Ces dernières offrent une alternative robuste aux données d'accidents, souvent rares ou incomplètes, et permettent une gestion proactive des risques. Les méthodes de prévision à court terme du trafic et les indicateurs de conflits tels que le TTC ou le TPA illustrent un tournant vers des approches plus dynamiques et adaptatives.

En somme, la littérature examinée met en évidence une évolution vers des systèmes intelligents et connectés, où les données collectées en temps réel, combinées à des algorithmes d'IA, ouvrent la voie à une gestion plus fine et préventive du trafic routier. Ce cadre théorique justifie pleinement la démarche innovante proposée dans le reste du travail, en répondant à des lacunes identifiées tant sur le plan technologique que méthodologique.

Chapitre 3 - Analyse du trafic routier à Trois-Rivières

Ce chapitre constitue une étape essentielle dans la compréhension des données collectées grâce aux capteurs *LiDAR* installés aux intersections de la Ville de Trois-Rivières. Avant toute modélisation ou application de techniques d'intelligence artificielle, il est fondamental de bien connaître la nature, la structure, la qualité et la distribution des données disponibles.

L'analyse exploratoire présentée ici permet ainsi : de dresser un portrait global des jeux de données (volume d'usagers, conflits, violations de feux rouges); d'identifier les variables clés et leur comportement statistique; de détecter d'éventuelles valeurs aberrantes ou biais de collecte; d'orienter les choix méthodologiques ultérieurs.

Nous introduisons d'abord les concepts théoriques nécessaires à l'interprétation des séries temporelles et des phénomènes stochastiques liés au trafic. Nous poursuivons avec une présentation détaillée des données disponibles, suivie d'une exploration visuelle et statistique des indicateurs mesurés.

3.1 Rappels théoriques

Ce chapitre explore les fondements de la démarche d'analyse au sein du projet, les justifications théoriques. Il vise à poser les bases nécessaires à la compréhension des choix méthodologiques, des outils utilisés et des résultats exploratoires obtenus pour orienter les options. On y trouve notamment une présentation du contexte d'analyse, les objectifs poursuivis, ainsi que les étapes clés du processus. Ce chapitre sert ainsi de pivot entre la conceptualisation du problème et son traitement concret.

3.1.1 Procédés stochastiques

Les procédés stochastiques sont des processus mathématiques qui évoluent dans le temps de manière aléatoire. Ils sont utilisés pour modéliser des phénomènes incertains en finance, en ingénierie, en physique, en biologie, et en d'autres domaines. Un processus stochastique est défini comme un ensemble de variables aléatoires ordonnées dans le temps et définies sur un ensemble de points pouvant être discrets ou continus.

La variable aléatoire à l'instant t est notée $X(t)$ lorsque le temps est continu, ou X_n lorsque le temps est discret. Un processus stochastique continu est représenté par l'ensemble $\{X(t); -\infty < t < +\infty\}$, tandis qu'un processus stochastique discret est représenté par l'ensemble $\{X_n; n = \ldots, -2, -1, 0, 1, 2, \ldots\}$. Les processus stochastiques sont largement utilisés en modélisation du trafic routier, notamment pour analyser les flux de véhicules, les temps d'attente, et les mesures de sécurité comme le TPA (Temps Post-Accrochage).

3.1.2 Séries temporelles

Les séries temporelles sont des suites de données collectées et ordonnées chronologiquement. Elles sont couramment utilisées pour analyser, modéliser et prédire des phénomènes évoluant dans le temps, comme les variations météorologiques, les fluctuations financières ou les débits de trafic routier. Une série temporelle est généralement représentée par une suite de valeurs : $\{y_n\}_{n=1}^{N}$ où y_n est la valeur observée à l'instant n et N est le nombre total d'observations.

Les composantes typiques d'une série temporelle sont la tendance ou l'évolution générale à long terme, la saisonnalité ou les motifs périodiques récurrents (ex. : variations quotidiennes ou annuelles), le cycle des fluctuations non périodiques et les variations aléatoires imprévisibles. Les modèles classiques des séries temporelles sont :

- le modèle AR (Auto Régressif) défini à l'équation (3.1) où la valeur actuelle dépend des valeurs passées.

$$y_n = \phi_1 y_{n-1} + \phi_2 y_{n-2} + ... + \phi_p y_{n-p} + \varepsilon_n \tag{3.1}$$

- le modèle MA (Moyenne mobile *Moving Average*), illustré par l'équation (3.2) où la valeur actuelle est basée sur les erreurs passées.

$$y_n = \varepsilon_n + \theta_1 \varepsilon_{n-1} + ... + \theta_q \varepsilon_{n-q} \tag{3.2}$$

- et le modèle ARMA (*Auto Regressive Moving Average*) qui combine les modèles AR et MA (3.3).

$$y_n = \varepsilon_n + \sum_{i=1}^{p} \varphi_i y_{n-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{3.3}$$

Les fonctions d'autocorrelation ACF (*Auto Correlation function*) et d'autocorrelation partielle PACF (*Partial Auto Correlation Function*) sont des outils essentiels pour analyser les séries temporelles, en particulier pour identifier les structures temporelles et choisir les bons modèles de prédiction comme les modèles AR, MA ou ARIMA (*Auto Regressive Integrated Moving Aerage*).

3.1.3 Fonction d'auto-corrélation

La fonction d'autocorrélation (ACF) mesure le degré de corrélation entre les observations d'une série temporelle à différents décalages temporels (ou lags). L'ACF à un décalage k est définie par (3.4):

$$\rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \tag{3.4}$$

où : y_t est la série temporelle, $Cov(y_t, y_{t-k})$ est la covariance entre les observations séparées par k périodes, $Var(y_t)$ est la variance de la série temporelle. Un $\rho_k = \pm 1$ signifie une forte corrélation (positive ou négative), alors que $\rho_k = 0$ veut dire qu'il n'y en a pas.

La fonction ACF permet d'identifier les composantes MA, de détecter la saisonnalité (pics périodiques dans l'ACF), d'évaluer la stationnarité de la série : une série stationnaire a généralement des valeurs d'ACF qui décroissent rapidement.

3.1.4 Fonction d'auto-corrélation partielle

La fonction d'autocorrélation partielle (PACF) mesure la corrélation entre les observations séparées par k périodes, en éliminant l'influence des lags intermédiaires. La PACF est définie comme la corrélation entre y_n et y_{n-k} après avoir retiré l'effet des termes intermédiaires y_{n-1}, y_{n-2}, ..., y_{n-k+1}. Une valeur élevée à un décalage k indique une relation directe significative à ce lag. Une valeur proche de zéro indique une absence de relation directe. La formule générale de la fonction est donnée par:

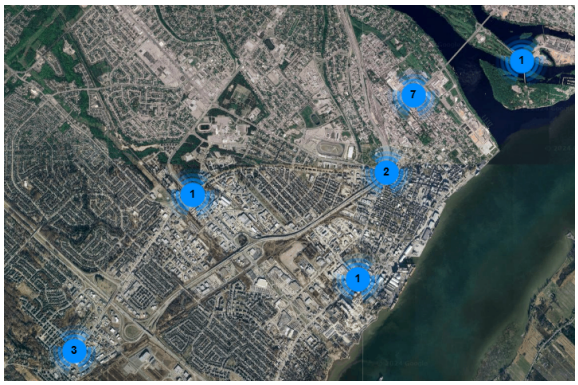
$$\phi_k = \text{corr}\,(y_n, y_{n-k} \mid y_{n-1}, y_{n-2}, \ldots, y_{n-k+1}) \tag{3.5}$$

En ce qui concerne la fonction PACF, elle sert à identifier les composantes AR et à déterminer l'ordre optimal p d'un modèle AR (en repérant le décalage à partir duquel la PACF devient non significative proche de zéro. La limite de signification est choisie de façon arbitraire.

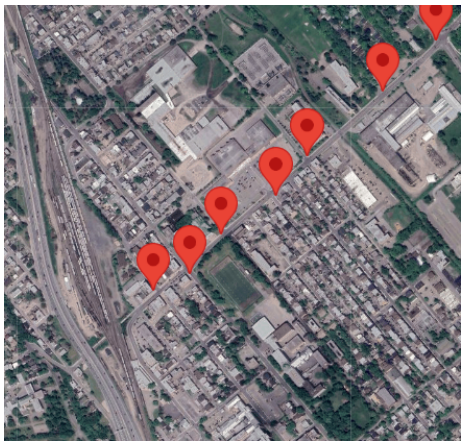
3.2 Présentation des données

Les données fournies par les capteurs LiDAR portent sur les conflits entre usagers, les violations de feux rouges et le compte de volume. Les nuages de points issus des capteurs prétraités permettent de détecter tout usager de la route, son mouvement, sa direction, entre autres, compter le volume sur une plage horaire, les TPA entre chacun, les violations de feux rouges. En effet, ces paramètres sont les principaux champs d'information fournis par la tierce partie partenaire *Blue City* propriétaire du matériel installé pour les mesures du trafic. Les installations portent sur une quinze (15) intersections dans des artères stratégiques telles qu'illustrées par la figure 3.1. La table 3-1 illustre ces quinze (15) intersections, leur code d'identification et les données disponibles, à savoir : les conflits, les violations de feu rouge et le volume d'usager.

Nous remarquons que les données de volume sont disponibles pour toutes les intersections. Il s'agit du nombre d'usagers selon les catégories : piétons, véhicules, bicyclettes, bus et camions. À chaque quinze minutes, le volume d'usagers motorisés présents à l'intersection est donné selon les directions d'arrivée et de destination (*NorthEast – NE*), et les traversées de gauche à droite en plus de la direction (*north left right – nlr*) pour le volume des usagers non motorisés (piétons et bicyclettes) sur support d'un fichier Excel (table 3-2). En ce qui concerne les conflits tels que montrés par la table 3-3, ils sont indiqués par : le TPA, qui est le temps entre le passage d'un premier et d'un second usager à un point géographiquement identifié (latitude et longitude), et la vitesse qui est impliquée. Des caractéristiques supplémentaires intéressantes sont fournies qui distinguent les conflits. Celles-ci sont : les types d'usagers impliqués, leurs mouvements respectifs, la date, l'heure (précision à la seconde) sont disponibles et annoncent une possibilité de classification. En outre, nous avons jugé utiles de calculer le rapport TPA sur la vitesse et d'ajouter celui-ci au



(a) Position des capteurs LiDAR. Les chiffres dans les points bleus représentent le nombre d'intersection



(b) Sept capteurs LiDAR sur St-Maurice

Figure 3.1 Installation des LiDAR

jeu de données dans le but de donner une valeur et un niveau de sévérité théorique. Seules quatre (4) intersections manquent de données de conflits : **107003**, **107008**, **107016** et **107018**. Et les données de violations de feu rouge sont seulement obtenues aux intersections **107002** et **107006**. Elles comprennent l'usager (motorisé), la phase de la lumière, l'heure de la phase, la date, le temps d'entrée et de sortie, le mouvement, la couleur de la lumière et les différences d'entrée et de sortie par rapport à l'heure de la phase (table 3-4).

Il est important de préciser que les données de conflits et de violations de feux rouges sont événementielles et renferment un caractère aléatoire. Par contre, les données de volume sont temporelles avec une période et une fréquence déjà établies. Cette précision est pertinente dans la mesure où ces particularités influencent à la fois la manière de les présenter dans les sections suivantes et les techniques appliquées pour chaque type de jeu de données.

3.2.1 Les conflits aux intersections entre usagers de la route

Concernant les données de conflits, celles qui servent directement à la sécurité routière, la mesure essentielle est le TPA dont les valeurs sont comprises entre 0 et 10 secondes. Le

Tableau 3-1 Données disponibles pour les intersections

Identification	Conflits	Violation Feu Rouge	Volume
107002	*X*	*X*	*X*
107003			*X*
107005	*X*		*X*
107006	*X*	*X*	*X*
107007	*X*		*X*
107008			*X*
107010	*X*		*X*
107011	*X*		*X*
107012	*X*		*X*
107013	*X*		*X*
107014	*X*		*X*
107015	*X*		*X*
107016			*X*
107017	*X*		*X*
107018			*X*

Tableau 3-2 Exemple de données de volume de piétons

INDEX	TEMPS	nlr	nrl	slr	srl	wlr	wrl	elr	erl
0	'2021-05-04 10:00'	10	8	2	31	1	1	15	24
1	'2021-05-04 10:15'	6	13	11	40	4	3	19	21
2	'2021-05-04 10:30'	12	5	7	32	0	12	11	10
3	'2021-05-04 10:45'	23	18	0	21	2	7	8	8
4	'2021-05-04 11:00'	14	10	6	16	3	7	5	14
5	'2021-05-04 11:15'	3	5	4	10	5	6	2	12

TPA est une mesure temporelle utilisée pour évaluer le niveau de risque lors d'un conflit entre deux usagers de la route. Il correspond à l'intervalle de temps entre l'instant t_1, où le premier usager quitte le point de conflit, et l'instant t_2, où le second usager atteint ce même point, comme illustré à la figure 3.2. Plus la valeur du TPA est faible, plus le risque de collision est élevé. Une valeur de TPA inférieure à zéro indique que les deux usagers ont occupé le point de conflit simultanément, ce qui correspond à une situation de collision [1].

Tableau 3-3 Exemple de données de conflits. L'étoile * indique que le mouvement ne s'applique et concernent les vélos et les piétons

TPA	VITESSE	TYPE1	TYPE2	MOUV1	MOUV2	DATE	LONG	LAT	SEVERITÉ
2.4	**30.5**	**Car**	**Bus**	**WBLT**	**SBST**	**2023-03-04 18:50**	**-72.35**	**42.34**	**Slight**
1.0	**42.2**	**Car**	**Car**	**NBST**	**SBLT**	**2023-03-04 20:02**	**-72.35**	**42.34**	**Serieux**
0.5	**15.1**	**Car**	**Piéton**	**WBST**	**'*'**	**2023-03-04 21:45**	**-72.35**	**42.34**	**Slight**
1.2	**60.2**	**Camion**	**Car**	**NBRT**	**EBST**	**2023-03-05 01:10**	**-72.35**	**42.34**	**Serieux**
7.8	**28.1**	**Bicycle**	**Car**	**'*'**	**SBST**	**2023-03-05 08:20**	**-72.35**	**42.34**	**Normal**
3.1	**23.0**	**Car**	**Car**	**SBST**	**EBST**	**2023-03-05 10:41**	**-72.35**	**42.34**	**Slight**

Tableau 3-4 Exemple de violation de feu rouge

Type	Phase	Début Phase	Temps d'entrée	Temps de sortie	Mouv.	État du feu	Diff. entrée	Diff. sortie
'Car'	**4**	**'17:31:46.4'**	**'17:31:46.2'**	**'17:31:48.7'**	**'SBLT'**	**Rouge**	**-0.2**	**2.3**
'Bus'	**4**	**'18:01:06.5'**	**'18:01:06.2'**	**'18:01:06.9'**	**'NBST'**	**Yellow**	**-0.3**	**0.4**

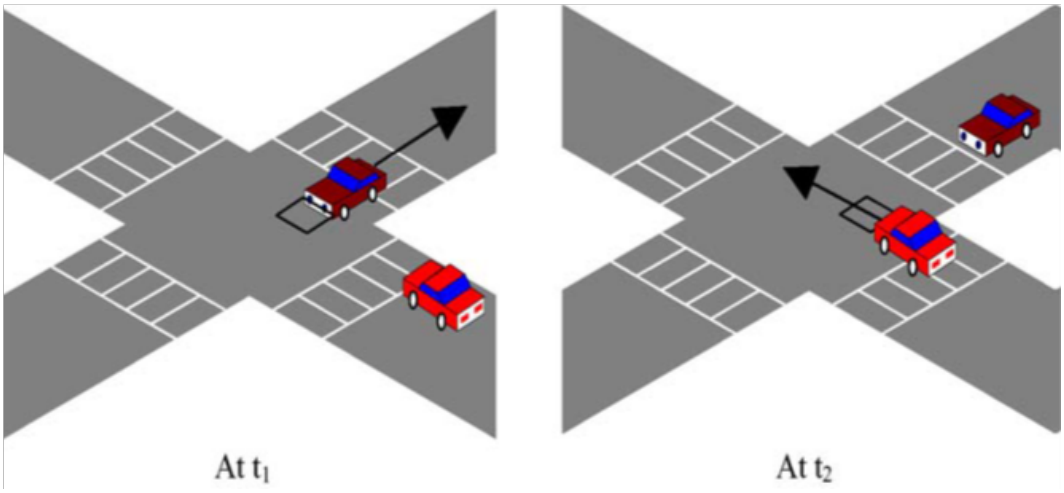


Figure 3.2 Illustration de la mesure de TPA [1]

Mais ce temps à lui seul ne saurait définir le niveau de sévérité du conflit ; par conséquent le risque d'un accident par exemple : un conflit peut survenir quand une voiture s'arrête pour laisser un piéton traverser la chaussée et reprend son chemin quelques secondes plus tard. Alors que dans la plupart de la littérature scientifique, ce TPA est utilisé pour définir si un potentiel de conflit est dangereux ou non selon que la valeur en secondes est en dessous d'un seuil. Nous allons intégrer la vitesse impliquée, une deuxième mesure qui joue un rôle intéressant, pour avoir une manière plus représentative du risque. Celle-ci renseigne un peu plus sur la distance qui sépare le deuxième usager du point conflictuel. Car il est connu que plus la vitesse est grande, plus il faut une grande distance pour réduire le risque et la sévérité d'un accident potentiel [13]. Elle vient bonifier notre appréciation de la valeur de TPA. En conséquence, une colonne de niveau de sévérité du conflit est donc rajoutée dans nos champs d'informations utiles tel qu'indiqué sur la Table 3-3.

La table 3-5 montre les différentes informations obtenues pour les conflits. Les deux mesures TPA et la vitesse sont les seules caractéristiques du jeu de données de conflit en rapport direct avec l'indice de performance de sécurité (une intégration d'avec les données de compte de volume est à considérer quoique). Toutefois, des informations telles que les usagers impliqués, les directions (de mouvements, d'origines et de destinations), le jour et le mois, les heures de la journée sont disponibles et annoncent une possible classification du couple de données vitesse-TPA.

Aussi, les points de conflit peuvent se diviser en trois catégories selon la figure 3.3 : les conflits divergents (*diverging*), convergents (*merging*) et de croisement (*crossing*) selon [2], ce qui nous rajoute une autre colonne de catégorie du type de conflit. Ces catégories peuvent être pertinentes car les potentiels accidents qui en suivraient seraient de niveaux différents en termes de gravité.

Tableau 3-5 Informations sur les conflits

Caractéristique	**Description**	**Type**	**Intervalle**
TPA	Temps séparant deux usagers du même point de conflit	Continue	0 - 10 secondes
Vitesse	Différence de vitesse entre les deux usagers	Continue	> 10 km/h
Premier Usager		Catégorie	Véhicule, piétons, bus, bicyclette, motocyclette, camions
Deuxième Usager		Catégorie	Véhicule, piétons, bus, bicyclette, motocyclette, camions
Mouvement 1er Usager	Directions d'arrivée et de destination	Catégorie	WBST, WBLT, WBRT, EBST, EBLT, EBRT, NBST, NBLT, NBRT, SBST, SBLT, SBRT
Mouvement 2e Usager	Directions d'arrivée et de destination	Catégorie	WBST, WBLT, WBRT, EBST, EBLT, EBRT, NBST, NBLT, NBRT, SBST, SBLT, SBRT
Date et heure	Date et heure de l'événement	Date	

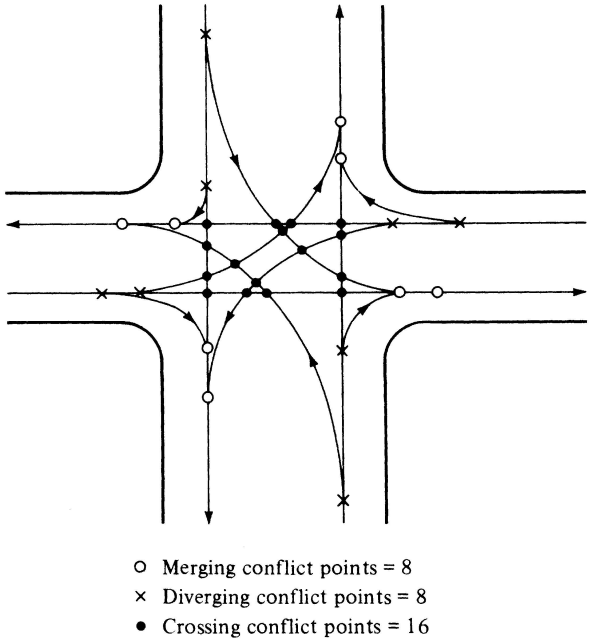


Figure 3.3 Type de conflits possibles dans une intersection de quatre directions d'arrivées [2]

3.2.2 Le volume d'usagers

Le volume d'usagers permet essentiellement d'avoir une idée sur la fluidité de l'intersection, sa capacité et quels axes sont plus ou moins surchargés. Une corrélation du volume et des conflits existe bel et bien. En effet, plus nous avons de monde, plus les conflits existent du fait de l'interaction continue des usagers de la route. Les données des volumes sont temporelles et il est intéressant d'utiliser des modèles prédictifs comme un filtre linéaire RLS ou le filtre de Kalman pour fournir des informations sur l'activité routière.

3.2.3 Les violations de feux rouges

Elles décryptent parfaitement le niveau de risque de sécurité d'une intersection lié au facteur humain. Elles sont pertinentes dans l'évaluation du risque. Toutefois, elles ne

sont disponibles que dans deux intersections, ce qui nous limite dans la possibilité de les combiner avec les autres données. Par conséquent, les travaux décrits au chapitre 4 ne portent pas sur les violations de feux rouges.

3.3 Exploration de données

Dans cette section, une description visuelle des données par des graphiques classiques oriente notre réflexion, explique et justifie notre approche. De l'histogramme pour l'affichage de chaque variable ou d'un nuage de points quand il s'agit d'un jeu de données avec plusieurs caractéristiques [19], sans oublier des matrices de nuage de points ou des distributions marginales sans exhaustivité, tout graphique serait utile pour mieux comprendre les données. La fréquence d'occurrence de chaque caractéristique des données de conflits et de volume d'usager sur chaque direction de chaque intersection nous intéresse.

Dans cette partie, nous évaluons dans la globalité des données les valeurs modales, les fonctions de distributions statistiques des observations pour mieux exploiter les ressources. Nous allons utiliser des combinaisons de champs d'informations disponibles aussi bien pour les conflits que pour les comptes de volumes. L'exploration vise à fournir des résultats quantitatifs après association et prétraitement.

3.3.1 Les conflits aux intersections entre usagers de la route

Les deux caractéristiques significatives de notre jeu de données des conflits sont le TPA et la vitesse, toutes les autres ne sont qu'une catégorisation selon les usagers, leur mouvement et la sévérité théorique. Nous justifions ce choix par l'hypothèse que les deux caractéristiques sont suffisantes pour établir le niveau de risque du conflit.

Densité de probabilité [19]

À partir de la définition de la densité de probabilité, si la variable aléatoire X a une densité f, définie par (3.6) :

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P\left(x - h < X < x + h\right) \tag{3.6}$$

quel que soit h, un estimateur de $P(x-h<X<x+h)$ est la proportion d'observations X_1, X_2,..., X_n comprises dans l'intervalle $[x-h, x+h]$ donné par (3.7) :

$$\hat{f}(x) = \frac{1}{2hn} N \tag{3.7}$$

où N est le nombre de fois $X_i \in [x-h, x+h]$, n étant le nombre d'observatons total de X_i.

Malheureusement, cet estimateur n'est pas une fonction continue et ne satisfait pas à la condition d'une estimation de densité. Cependant, une fonction d'estimateur de noyau est déduite de celle-ci et est donnée par (3.8) :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.8}$$

avec K une fonction qui peut être triangulaire, rectangulaire ou gaussienne.

La condition de densité donnée par (3.9):

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{3.9}$$

doit être entièrement satisfaite. Cette fonction de densité est généralement symétrique comme nous aurons à le constater. Car, pour une meilleure appréciation, nous avons utilisé les densités de probabilité gaussiennes des histogrammes des deux variables pour voir leur distribution.

Les figures 3.4 et 3.5 nous renseignent sur les distributions du TPA et de la vitesse. Globalement, pour toutes les intersections, les TPA les plus fréquents sont supérieurs à 2.5 secondes. En dessous de ce seuil, la fréquence est faible. Toutefois, nous remarquons que pour les intersections **107010** et **107014**, le seuil des fréquences élevées est doublé et est noté environ 5 secondes. Aussi, il faut souligner que les TPA en deçà de 2 secondes sont plus fréquents pour les intersections **107015** et **107012**. On peut aisément affirmer que ces deux dernières intersections présentent plus de risque d'avoir un TPA égal à ou inférieur à zéro (ce qui signifie un accident) et sont de bons candidats à évaluer. Excepté les intersections **107010** et **107014**, la fonction de densité est presque rectangulaire pour toutes les autres.

Généralement, les données de vitesse des intersections suivent une loi de distribution normale. Les vitesses sont aussi globalement dans des intervalles convenables. On voit facilement que la limite de vitesse est respectée à l'approche des intersections. Tout aussi pareil qu'au TPA, les intersections **107010** et **107014** ont des fréquences importantes aux petites vitesses et aucune au-delà de 40 km/h. Pour avoir des événements de vitesses approchant le seuil de 60 km/h, il faut regarder les intersections **107005**, **107012** et **107017**.

Les observations sont confirmées lorsqu'on regarde les distributions de densités de probabilités ensemble sur la figure 3.6. En outre, la figure 3.7 est intéressante dans la mesure où elle illustre comment les caractéristiques se combinent. Elle nous permet d'affirmer que dans la majeure partie du temps, les conflits sont à des niveaux acceptables (TPA > 2 s et

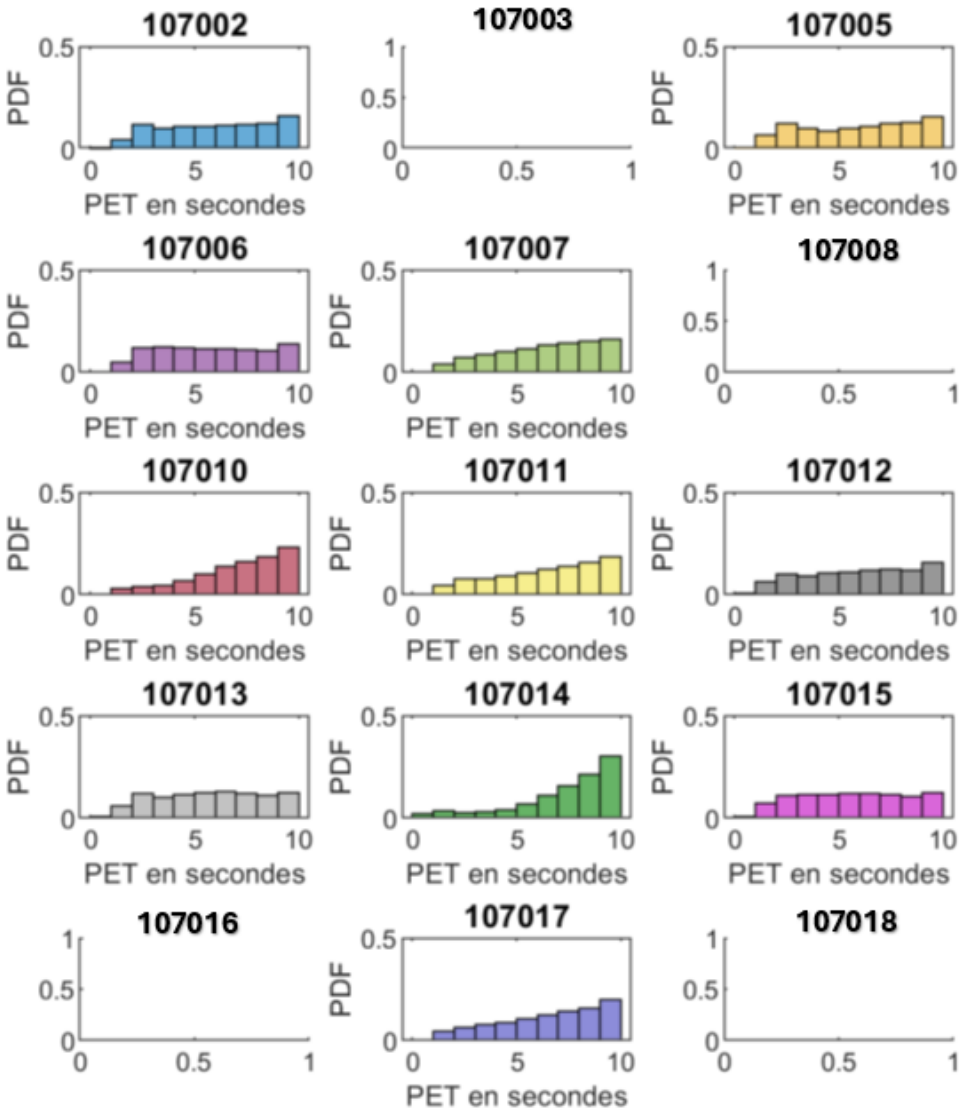


Figure 3.4 Histogramme des TPA de chaque intersection (ID)

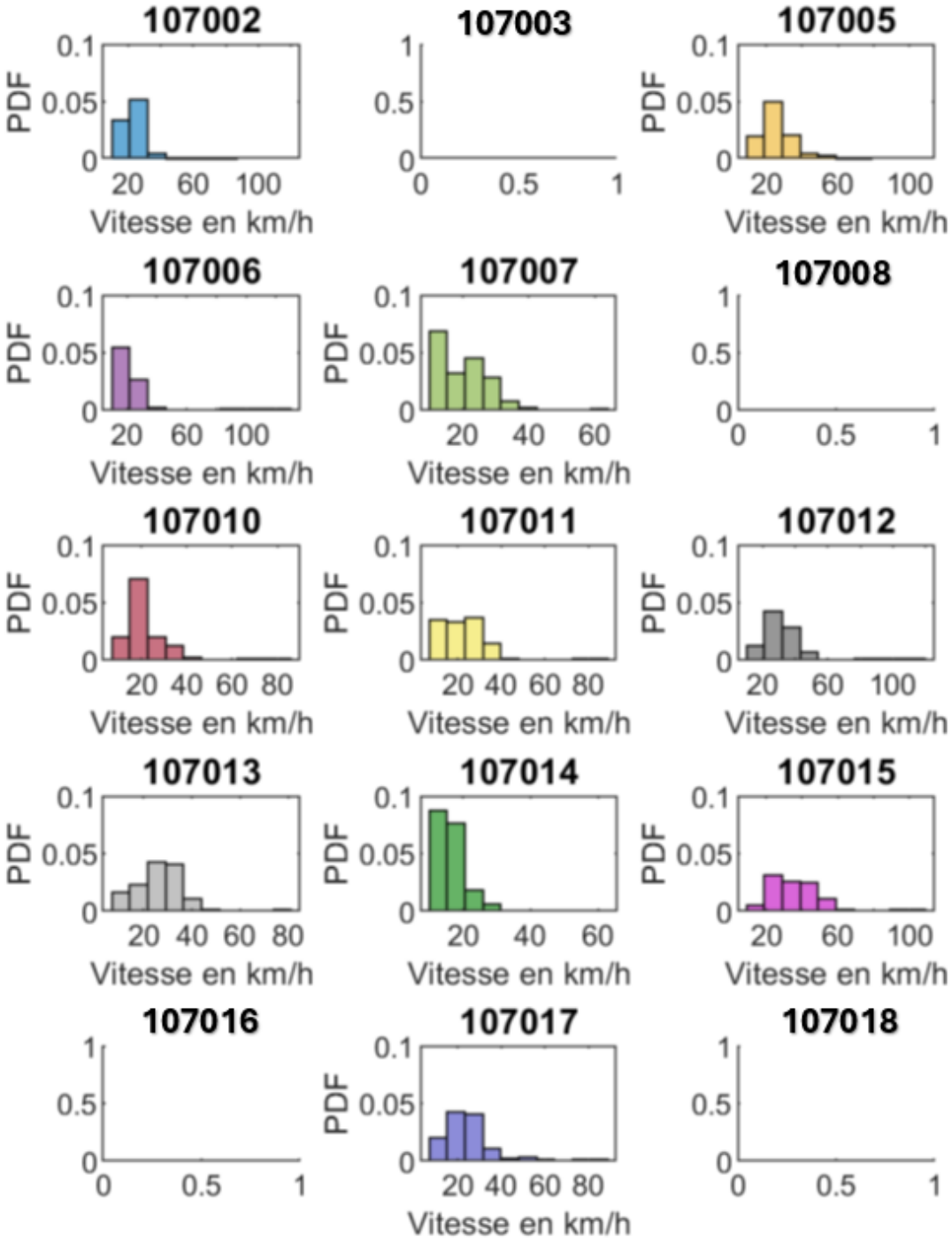


Figure 3.5 Histogramme des vitesses de chaque intersection (ID)

Vitesse < 40 km/h). Étant donné que les situations dangereuses ne surviennent que très rarement, nous pouvons déjà orienter notre approche vers l'étude de ces raretés.

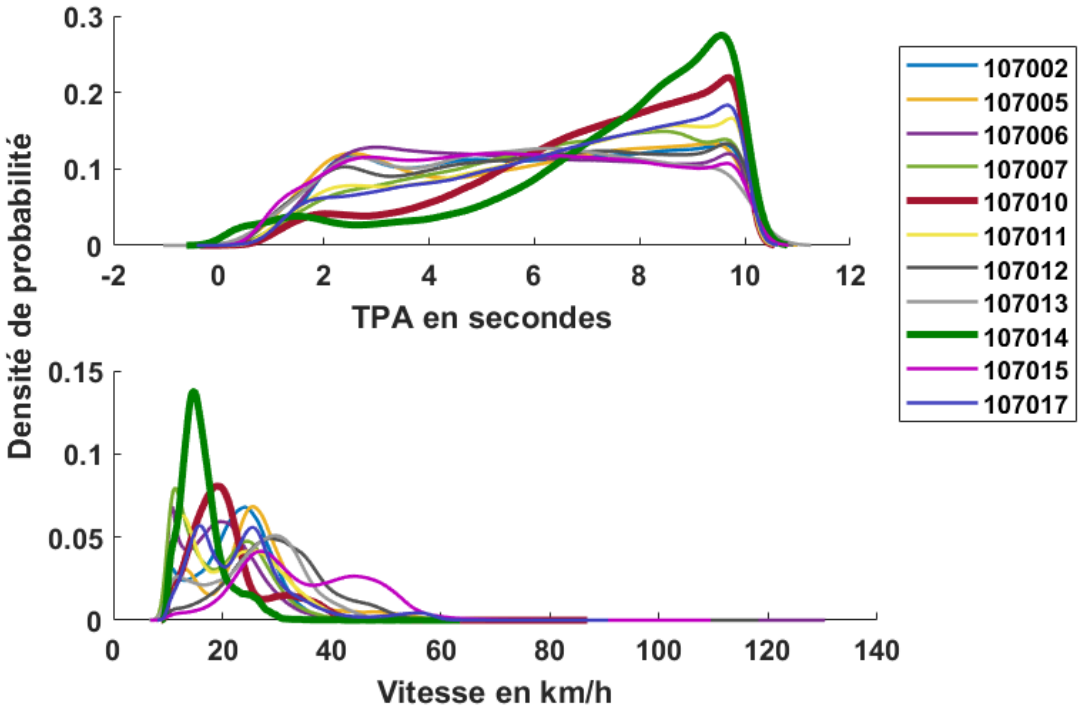


Figure 3.6 Densité de probabilité des données de conflits

Plus loin dans l'interprétation, le diagramme à moustache des TPA comme indiqué aux figures 3.8 et 3.9 décryptent mieux les similitudes entre intersections et instinctivement nous mettent en lumière les asymétries des distributions. On voit que pour la vitesse, les données aberrantes sont plus fréquentes, justifiées par le fait que cette caractéristique n'a pas de limite maximale. Les diagrammes confirment que la quasi-totalité des données de conflits (estimable à environ 90 %) sont dans des intervalles "normaux" car d'après ces figures, 75 % des vitesses et TPA sont respectivement inférieures à 40 km/h et comprises entre 4 à 8,5 secondes. Nous pouvons remarquer et affirmer avec certitude que l'intersection **107014** ne présente aucune inquiétude du point de vue sécuritaire.

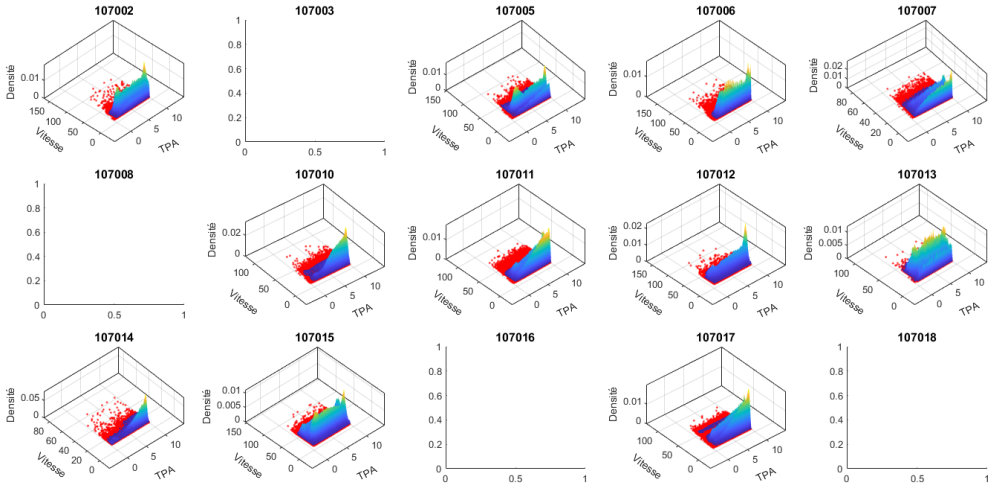


Figure 3.7 Distribution jointe des données TPA et Vitesse

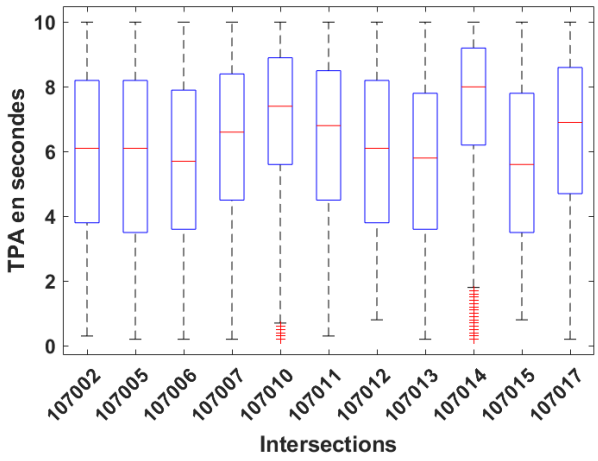


Figure 3.8 Diagramme à moustache des données de TPA

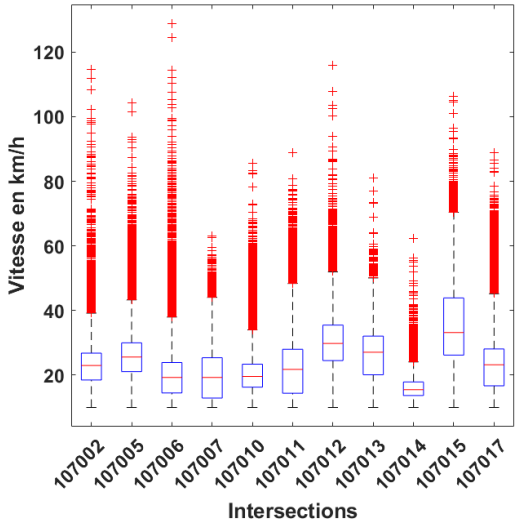


Figure 3.9 Diagramme à moustache des données de vitesse

Toutefois, les diagrammes donnent des intuitions sur les données aberrantes. Celles-ci doivent être examinées pour voir si elles sont vraiment à considérer comme telles ou peuvent contenir des renseignements pertinents pour l'analyse de la sécurité. Dans l'objectif d'étudier l'aspect sécuritaire de façon plus rigoureuse, une analyse avec des méthodes statistiques plus poussées est nécessaire avant le calcul du risque. Celles-ci sont décrites et présentées dans le chapitre 4.

Nous avons vu comment se distribuent le couple de vitesse et de TPA selon une classe d'information comme le type d'usager, le mouvement et l'année. Par la suite, l'exploration des données vise à voir si des catégories de mouvement, d'usagers ou des combinaisons de type de ces deux se distinguent dans les caractéristiques.

Selon la catégorie du premier et du second : motorisé ou non

Les catégories de *LEADING* et *FOLLOWNG CLASS TYPE* (premier et deuxième usager qui traversent le point de conflit) sont réorganisées en groupes: *voiture*, *bus*, *camions* et

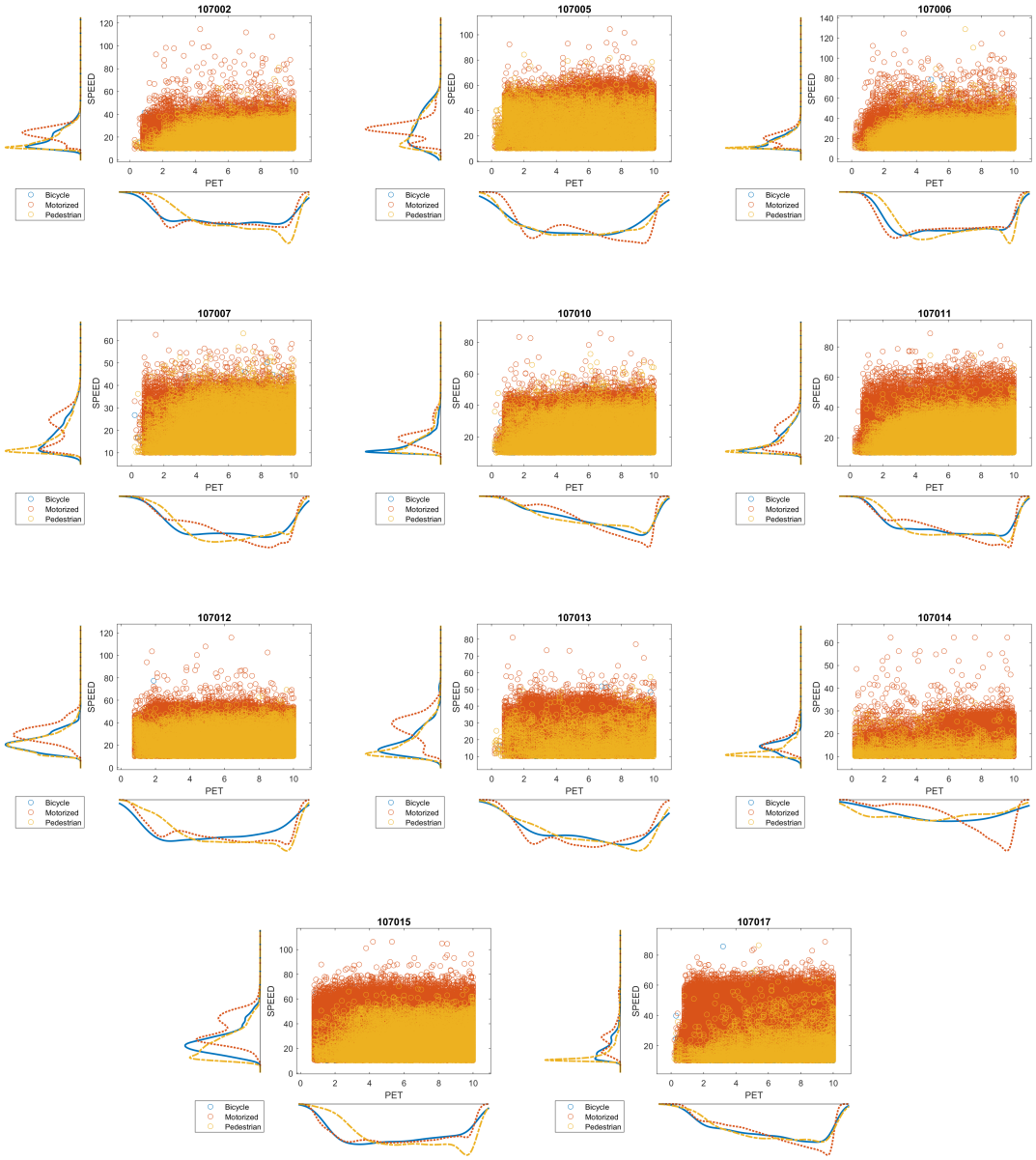


Figure 3.10 Distribution de vitesse et TPA selon le type d'usager : motorisé ou non du premier usager

motocyclette, pour le groupe motorisés et les types *piétons*, *bicyclette* sont gardés comme tels. En observant la distribution de la vitesse et TPA, aucun groupe ne se distingue de l'autre (figure 3.10).

Selon les catégories de mouvements du premier et du second

Les catégories de *LEADING* et *FOLLOWING MOVMENT* (mouvement du premier et du deuxième usager qui traversent le point de conflit) sont : *Direction Nord virage droite, gauche ou tout droit*, *Direction Sud virage droite, gauche ou tout droit*, *Direction Est virage droite, gauche ou tout droit*, *Direction Ouest virage droite, gauche ou tout droit*, *Traversée Piétons et bicyclette, ou passage motocyclette*. Une simplification est faite pour réduire le nombre de groupes distincts comme à la section précédente. Nous avons les groupes suivants : *Virage à droite*, *Virage à gauche* et *Tout droit*. Aucun groupe ne se distingue de l'autre (figure 3.11).

Selon le jour de la semaine

Les jours passent et se ressemblent pour toutes les intersections (figure 3.12).

Selon les années

Comme l'illustre la figure 3.13, les densités de TPA sont généralement pareilles dans les deux années où les données sont disponibles, sauf pour **107005** où nous remarquons deux modes (dont l'explication pourrait être les travaux sur l'intersection durant l'intervalle entre les deux périodes). Un pic de fréquence de TPA 2023 et un autre en 2022 respectivement supérieur et inférieur à 5 secondes. Le risque est plus important en 2022 qu'en 2023. Concernant la vitesse, cette assertion est plus que vraie. En 2023, il est parfaitement visible sur les graphiques que les densités de vitesse se tiennent en dessous de 20 km/h pour toutes

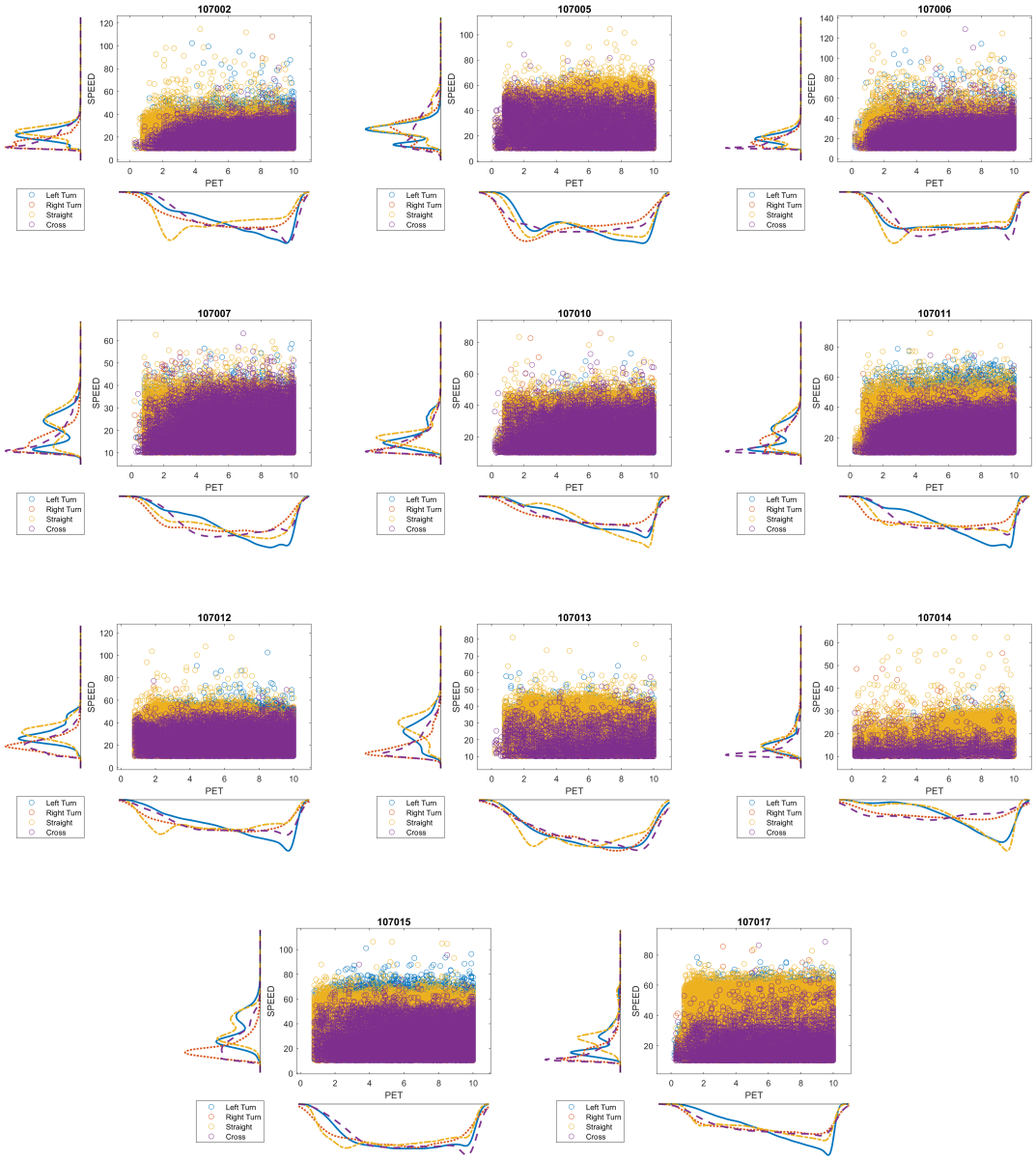


Figure 3.11 Distribution de vitesse et TPA selon le mouvement : virage à droite, à gauche ou tout droit du premier usager

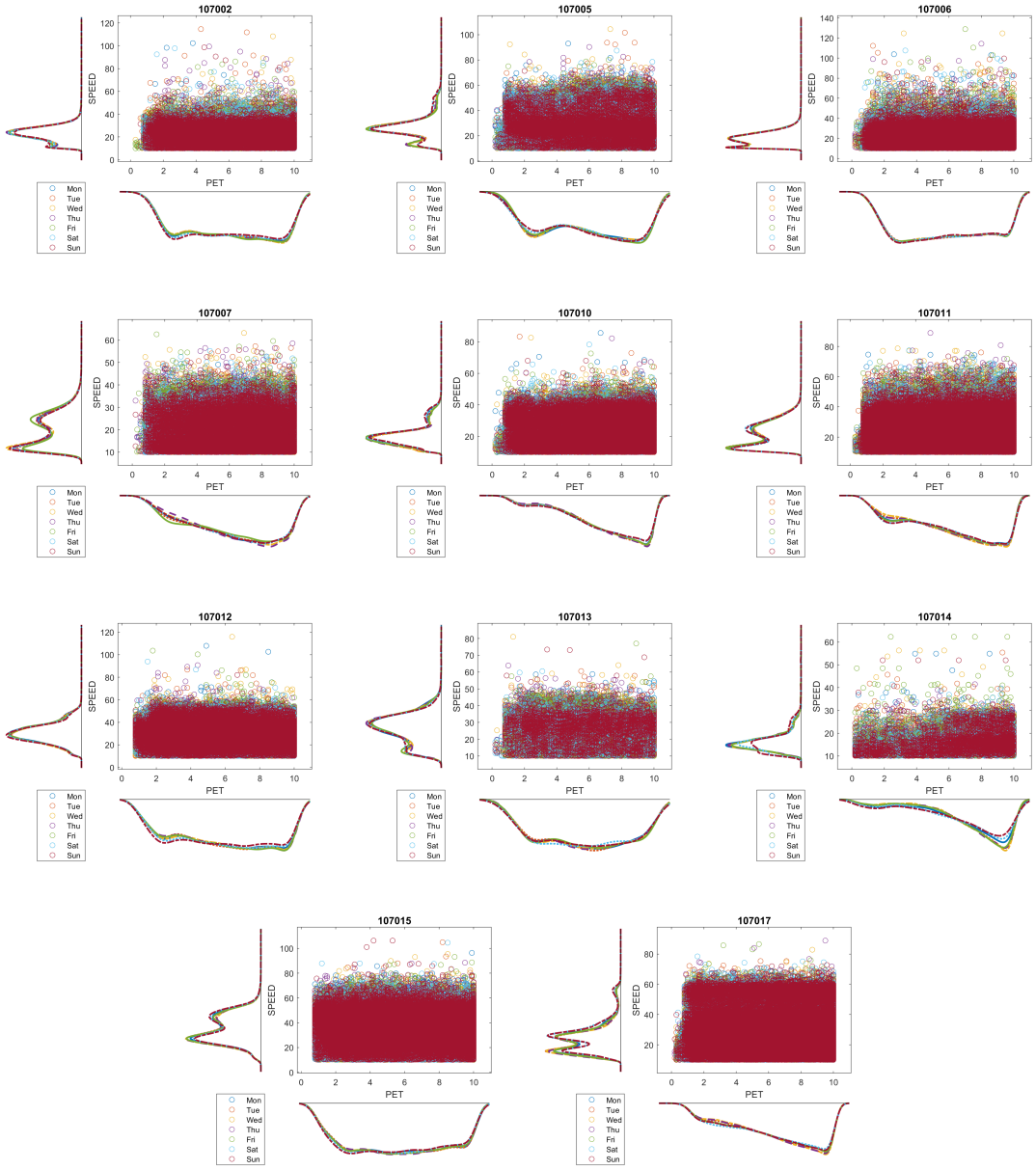


Figure 3.12 Distribution de vitesse et TPA selon le jour de la semaine

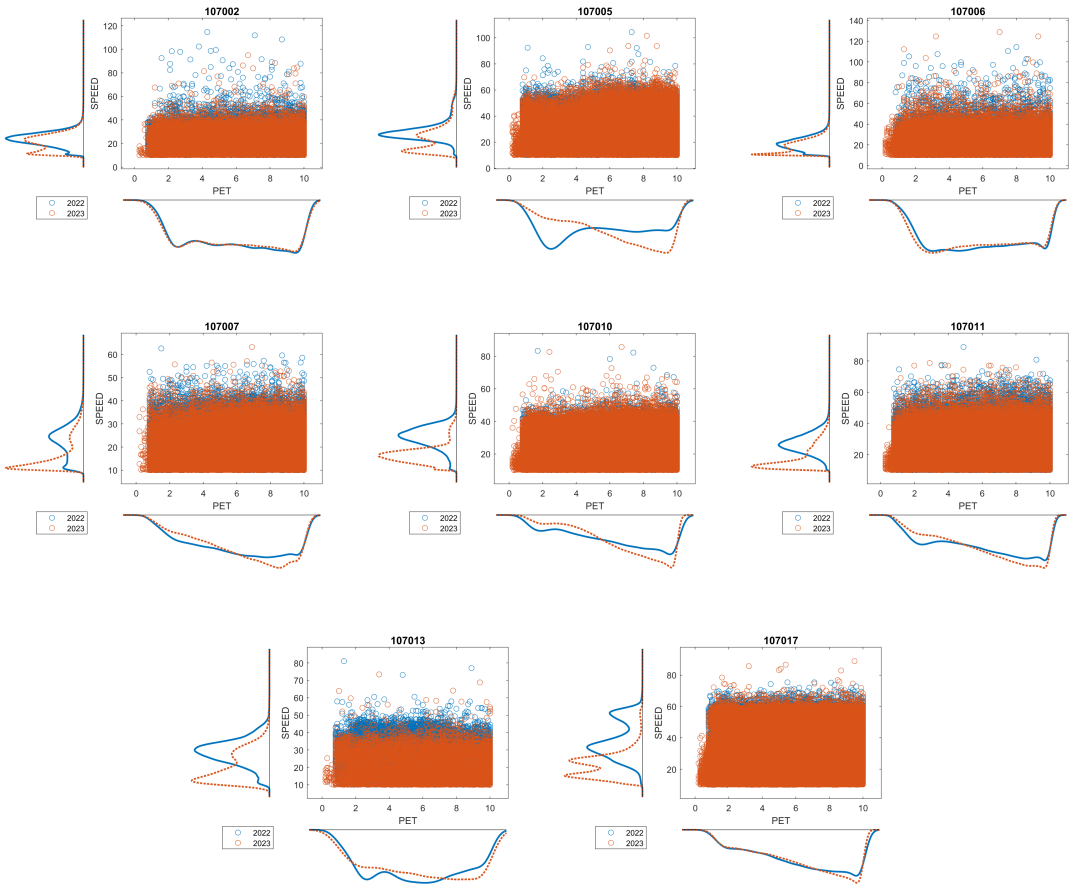


Figure 3.13 Distribution de vitesse et TPA selon les années

les intersections alors qu'en 2022 l'essentiel des vitesses se concentre en dessous de 30 km/h.

Recherche de patron

La répartition des données vitesse-TPA ne donne pas de séparations pertinentes entre catégories ou groupes de catégories. En l'état actuel, il ne serait pas nécessaire d'appliquer des techniques de regroupement (clustering) simples (par exemple le K-Moyenne) sur les nuages de points. Par contre, dans notre objectif final d'évaluer le niveau de risque, on utilise les techniques statistiques, particulièrement la théorie des valeurs extrêmes, pour prédire et classer les risques d'accidents.

3.3.2 Le volume des usagers de la route

Le jeu de données de volume compte chaque usager à toutes les quinze minutes des directions des axes routiers des intersections pour les usagers motorisés, alors que, pour les usagers non motorisés, nous avons des traversées de gauche à droite sur les axes. Le caractère principal des données de volume est l'aspect temporel.

Cela dit, les fonctions ACF et de PACF sont des outils essentiels pour analyser les séries temporelles, en particulier pour identifier les structures temporelles et choisir les bons modèles de prédiction comme les modèles AR, MA ou ARIMA. Mais, nous allons faire une analyse de corrélation entre les volumes des directions pour chaque intersection.

Analyse de la corrélation linéaire entre les directions

La corrélation renseigne sur la liaison entre les champs d'informations des données. Lorsque la corrélation est nulle entre deux champs, cela signifie que ceux-ci ne sont pas régis par la même dynamique, en d'autres termes, les informations ne réagissent pas

conjointement. Il est à distinguer que l'absence de corrélation ne signifie pas indépendance, une liaison non linéaire peut toujours subsister. Dans notre exploration des données temporelles comme celles des comptes de volume, pour chaque intersection, nous voulons voir comment les directions interagissent et la corrélation linéaire entre elles est un excellent moyen sans pour autant, à cet instant, modéliser les données.

Les corrélations entre directions sont présentées par les figures 3.14 à 3.18. Globalement pour toutes les intersections, la liaison linéaire entre les directions est quasiment insignifiante sauf les usagers de type véhicules qui sont plus concernés par les phases des feux de circulation. Concernant les piétons, il y a une très forte non-linéarité qui existe entre les directions de traversées, comme le montre la figure 3.14. Nous notons aussi sur celle-ci une forte corrélation des deux sens de l'axe Nord de l'intersection **107010**. Pour ce qui est des bicyclettes, à la figure 3.15, la corrélation est positive pour les traversées des axes Sud et Est (entre le sens gauche et droit) à l'intersection **107018**. En termes de corrélation, les usagers de véhicules sont les plus intéressants et nous observons que les axes vis-à-vis ont une liaison linéaire à cause du fait de la synchronisation des lumières. Ceci est particulièrement visible à l'intersection **107014** tel qu'indiqué à la figure 3.17. La couleur noires indiquent des données absentes où les cas ne s'appliquent pas.

Étude de l'auto-corrélation

La plupart des séries temporelles présentent une décroissance progressive des valeurs ACF, comme le montre la figure 3.19. Ce qui suggère une autocorrélation persistante et un comportement potentiellement non stationnaire. Cette décroissance lente peut indiquer une tendance ou une composante saisonnière.

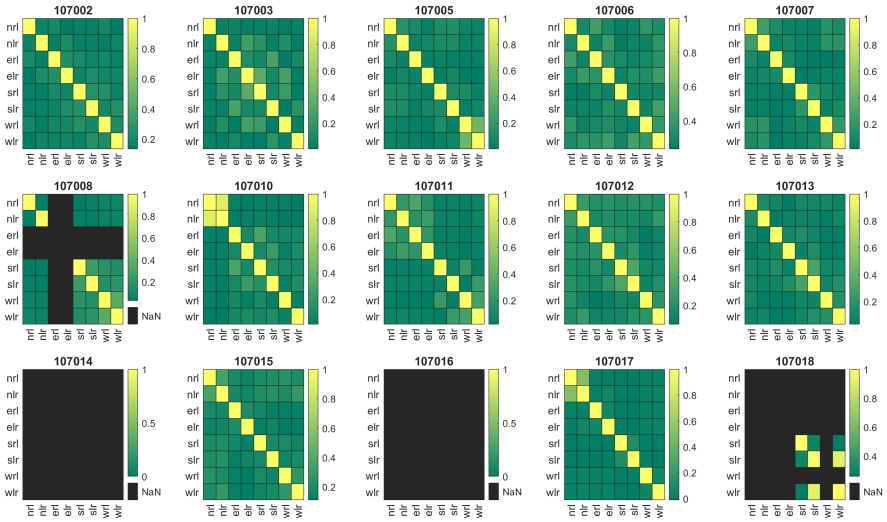


Figure 3.14 Corrélation de directions des piétons

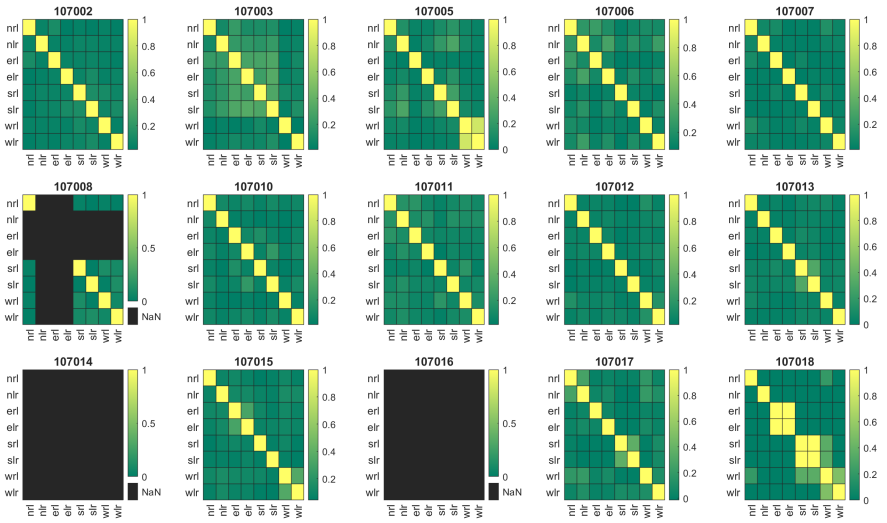


Figure 3.15 Corrélation de directions des bicyclettes

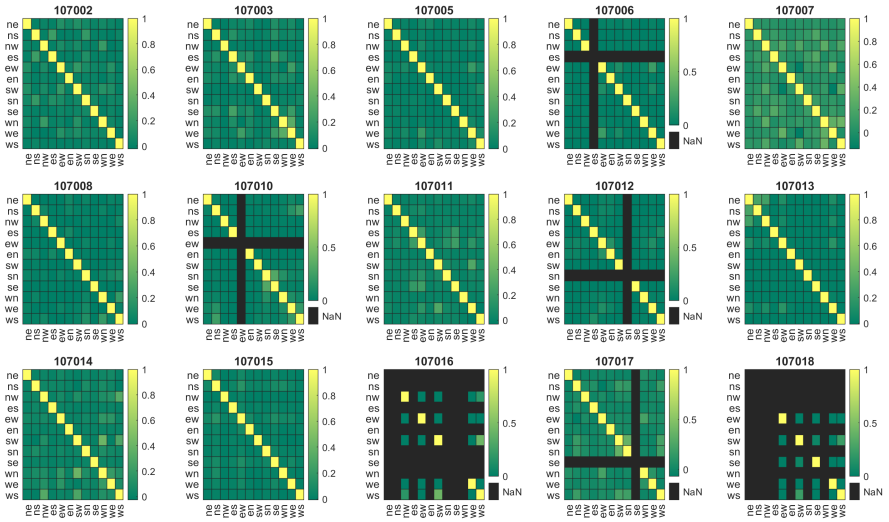


Figure 3.16 Corrélation de directions des bus

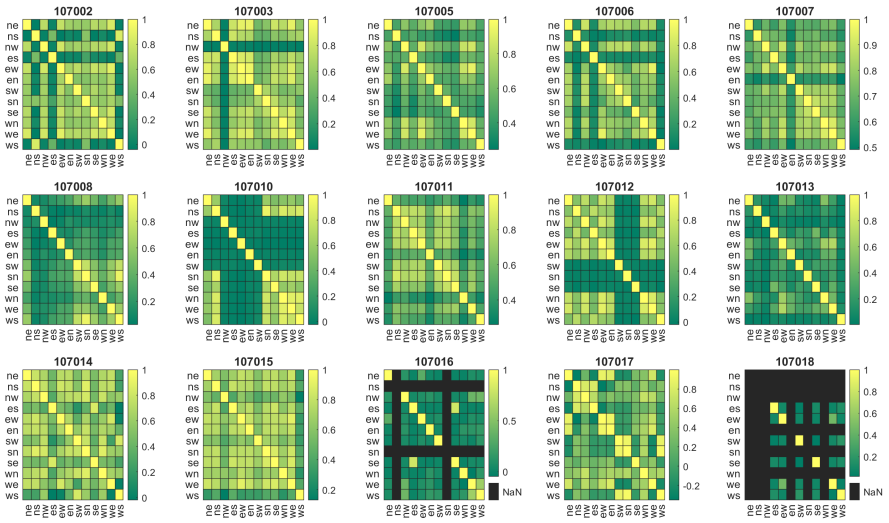


Figure 3.17 Corrélation de directions des véhicules

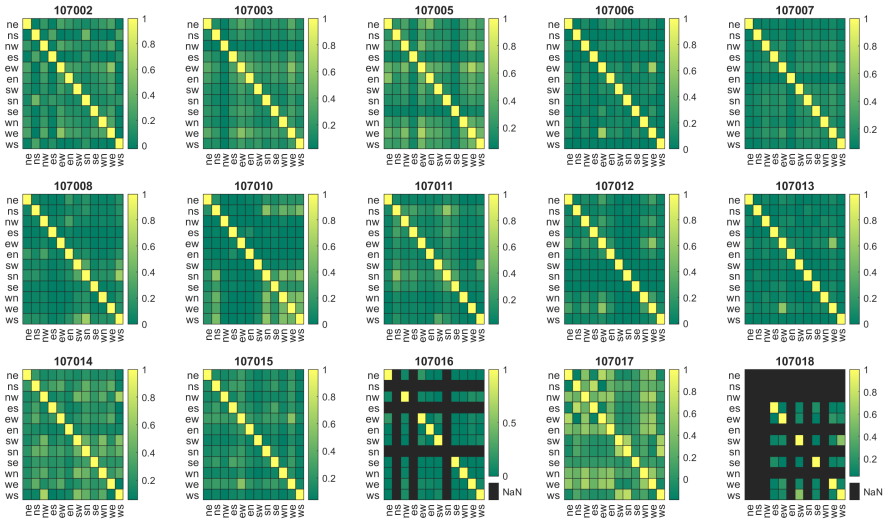


Figure 3.18 Corrélation de directions des camions

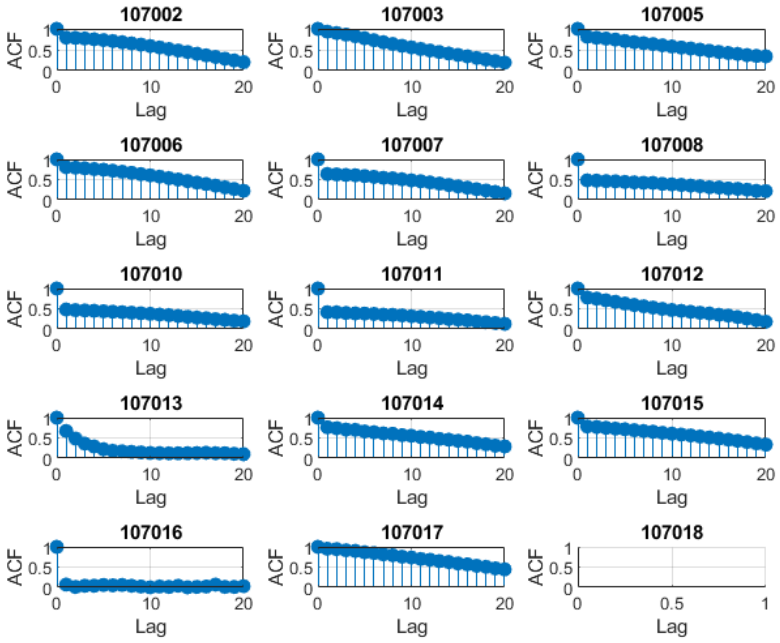


Figure 3.19 Auto-corrélation de la direction Nord-Est

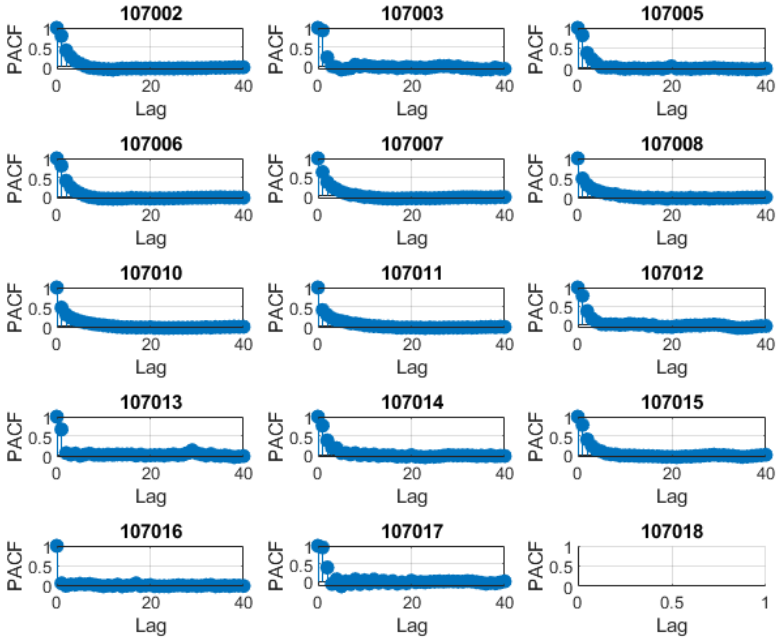


Figure 3.20 Auto-corrélation partielle de la direction Nord-Est

Une absence de corrélation significative (série potentiellement bruitée ou stationnaire sans structure temporelle apparente) est observée à l'intersection **107018**. La décroissance rapide de l'ACF suggère des séries temporelles plus stationnaires ou des processus autorégressifs d'ordre faible aux intersections **107013** et **107016**.

Étude de l'autocorrélation partielle PACF

À la figure 3.20, de nombreuses séries présentent une chute rapide après le premier ou le deuxième lag, ce qui indique un AR d'ordre faible. Pour les intersections **107013**, **107014** et **107016**, la PACF chute abruptement après le premier lag, ce qui renforce l'idée d'un processus AR(1).

Il n'y a aucune structure visible, cohérente avec le comportement en ACF de l'intersection **107018**. Aux intersections **107002, 107003 et 107005**, la décroissance lente de la PACF suggère un modèle AR plus complexe d'où la nécessité d'une différenciation.

Un modèle ARIMA avec différenciation peut être pertinent pour les séries avec une décroissance lente en ACF et une PACF qui chute rapidement. Un modèle ARIMA avec une forte composante autorégressive est recommandé pour les séries avec ACF et PACF décroissant lentement ensemble.

Optimisation ACF et PACF

L'ajustement optimal des paramètres du modèle ARIMA est crucial pour obtenir des prévisions précises. Les paramètres p et q sont tirés des fonctions ACF et PACF.

Propriétés des ACF/PACF

À la figure 3.21, l'ACF décroît lentement, ce qui est souvent un signe que les données ne sont pas stationnaires. Le PACF montre un comportement plus erratique, ce qui complique davantage l'identification correcte de l'ordre p. Nous identifions les ordres de $p = 8$ et $q = 25$.

Caractéristiques des données

Si les données sont bruitées, dynamiques ou présentent des changements de comportement rapide (comme c'est souvent le cas avec des données de trafic), un modèle plus adaptatif comme l'AEKF (filtre de Kalman étendu adaptatif) pourrait mieux suivre ces variations.

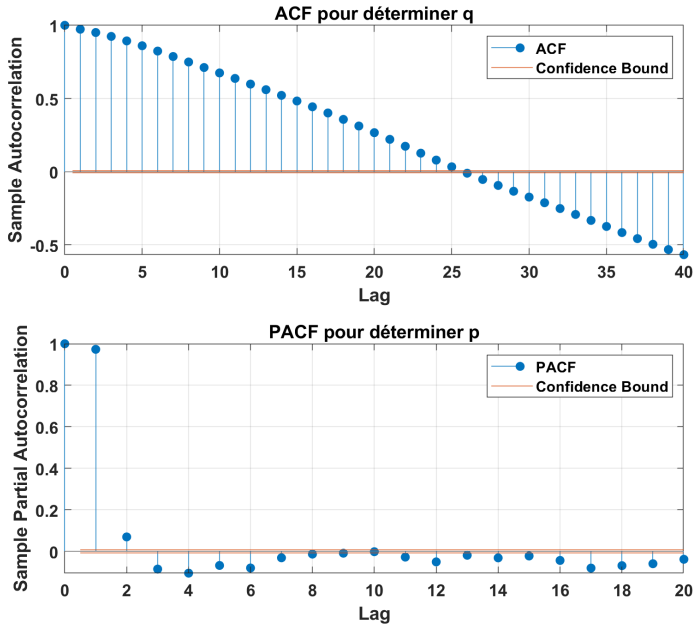


Figure 3.21 Optimisation de paramètres p et q d'un modèle ARMA. Exemple de l'intersection 107002

Modèle SARIMA

Étant donné que la série temporelle présente une saisonnalité quotidienne évidente, le modèle SARIMA (*Seasonnal Auto Regressive Integrated Moving Average*) est noté :

$$\text{SARIMA}(p,d,q) \times (P,D,Q,s) \tag{3.10}$$

où p,d,q représentent les composantes non saisonnières (ARIMA), P,D,Q,s désignent les composantes saisonnières avec s correspondant à la période de saisonnalité journalière. Ici la valeur de s est égale à 96.

L'équation mathématique d'un SARIMA du premier ordre, SARIMA(1,1,1), est formulée comme suit :

$$(1-\phi_1 B)(1-\Phi_1 B^s)(1-B)(1-B^s)y_t = (1+\theta_1 B)(1+\Theta_1 B^s)\varepsilon_t \tag{3.11}$$

où

- y_t est la valeur de la série à l'instant t, B est l'opérateur de délai, représentant l'équation de retard donnée comme suit :

$$B * y_t = y_{t-1} \tag{3.12}$$

- ϕ_1 est le coefficient autorégressif non saisonnier, Φ_1 est le coefficient saisonnier autorégressif, θ_t est le coefficient de la composante non saisonnière de la moyenne mobile, Θ_t est le coefficient saisonnal de la moyenne mobile, s est la période saisonnière, ε_t est le bruit blanc à l'instant t.

Analyse des aberrances

Il est courant d'être confronté à la présence d'aberrances dans les données dues à des erreurs de saisie. Mais dans un contexte comme le nôtre, où les données sont fournies par une tierce partie après un processus rigoureux de prétraitement, nous considérons que s'il y a aberrances, c'est dû au fait simplement qu'un élément du jeu s'éloigne de la moyenne. À cette étape de notre travail, il n'est pas fortuit de faire une analyse de l'inconsistance afin de voir la pertinence d'appliquer une technique de filtrage. Nous choisissons arbitrairement de considérer aberrante toute donnée représentée par X_n qui est supérieure à dix fois la valeur de la déviation standard (3.13).

$$X_n > 10\lambda \tag{3.13}$$

Tableau 3-6 Nombre de données aberrantes pour les piétons

	nlr	**nrl**	**elr**	**erl**	**slr**	**srl**	**wlr**	**wlr**
107002	15	10	8	11	12	14	3	9
107003	39	42	20	14	6	15	29	**161**
107005	24	29	**346**	**297**	76	29	88	**100**
107006	2	3	1	0	10	9	1	0
107007	2	5	22	29	22	34	6	6
107008	24	26	0	0	7	3	0	1
107010	73	69	24	27	19	8	71	68
107011	11	6	6	8	26	54	64	25
107012	1	1	12	8	11	10	17	9
107013	32	37	35	**132**	5	6	77	51
107014	0	0	0	0	0	0	0	0
107015	1	1	16	20	4	0	17	8
107016	0	0	0	0	0	0	0	0
107017	69	63	40	38	**141**	**288**	15	15
107018	0	0	0	0	7	4	0	1

les Tableaux 3-6 à 3-10 illustrent le nombre de données aberrantes pour toutes les directions de toutes les intersections et pour chaque type d'usager. Lorsqu'une intersection présente un grand volume d'usagers, il faut s'attendre à un grand nombre de données aberrantes. Ainsi, l'application de techniques de filtrage semble pertinente pour la plupart des intersections, mais par souci d'homogénéité, nous les appliquons à toutes les intersections.

Parmi ces techniques, nous avons le choix entre la fenêtre mobile, la moyenne mobile, la méthode Savitzky-Golay [20], entre autres. De façon arbitraire et pour des raisons de simplicité, nous choisissons la première méthode. La figure 3.22 montre les données brutes et filtrées par la méthode de la moyenne mobile sur une fenêtre de cinq pas de temps de l'intersection **107002** comme exemple.

Tableau 3-7 Nombre de données aberrantes pour les bicyclettes

	nlr	**nrl**	**elr**	**erl**	**slr**	**srl**	**wlr**	**wlr**
107002	33	19	**120**	62	8	45	21	30
107003	43	27	8	4	1	2	166	6
107005	18	34	51	**244**	91	75	81	**106**
107006	17	2	14	1	46	15	45	26
107007	48	17	**397**	**410**	**558**	**200**	68	**193**
107008	2	0	0	0	40	14	34	0
107010	25	48	10	25	53	16	48	9
107011	8	7	35	9	40	36	42	27
107012	3	18	14	9	21	72	15	5
107013	**228**	**211**	**288**	**120**	9	11	**214**	**314**
107014	0	0	0	0	0	0	0	0
107015	10	30	53	21	41	26	30	13
107016	0	0	0	0	0	0	0	0
107017	89	22	32	71	**113**	74	32	30
107018	3	2	1	1	1	1	8	2

Tableau 3-8 Nombre de données aberrantes pour les véhicules

Intersection	**ne**	**ns**	**nw**	**es**	**ew**	**en**	**sw**	**sn**	**se**	**wn**	**we**	**ws**
107002	0	90	0	53	0	0	1	67	4	3	0	84
107003	0	2	22	0	0	0	10	32	0	0	0	4
107005	20	18	1	1	0	82	0	43	25	0	0	2
107006	0	**113**	0	83	0	2	0	0	0	2	0	21
107007	6	0	0	0	0	6	0	0	0	0	0	0
107008	1	34	24	1	0	11	0	0	0	7	0	0
107010	4	0	32	**106**	**101**	**530**	57	0	17	0	0	0
107011	2	1	5	0	0	2	0	1	0	18	0	0
107012	0	14	0	14	0	0	**265**	**107**	**133**	0	0	4
107013	**103**	88	**373**	2	0	12	10	5	2	6	0	58
107014	1	0	2	**108**	0	2	0	0	97	13	0	0
107015	0	0	4	0	0	0	2	0	0	0	0	3
107016	97	0	6	1	11	24	8	0	6	12	5	2
107017	0	0	0	0	0	0	3	0	33	0	0	12
107018	0	0	0	22	0	0	4	0	14	0	0	2

Tableau 3-9 Nombre de données aberrantes pour les bus

Intersection	ne	ns	nw	es	ew	en	sw	sn	se	wn	we	ws
107002	**408**	65	2	10	1	96	**110**	**354**	50	3	0	9
107003	**162**	23	30	92	0	31	29	**197**	88	0	1	9
107005	83	**265**	51	28	6	72	**123**	**227**	**104**	**440**	2	18
107006	112	1	**574**	0	1	86	**418**	20	50	**232**	0	1
107007	**444**	**479**	0	**468**	0	**285**	0	33	98	0	0	0
107008	48	29	9	4	3	32	79	8	11	60	**156**	39
107010	**620**	63	3	1	0	3	**164**	88	**237**	40	0	2
107011	90	**141**	18	**370**	97	45	1	74	**118**	22	1	**428**
107012	39	41	**296**	66	0	**166**	1	0	7	88	34	24
107013	**148**	43	32	10	0	16	6	1	30	21	0	4
107014	**230**	88	39	85	73	**306**	32	51	**442**	0	**276**	87
107015	**149**	19	24	9	22	84	85	32	1	96	89	22
107016	0	0	6	0	1	0	10	0	0	0	10	17
107017	36	91	64	71	9	25	90	77	0	6	79	41
107018	0	0	0	0	3	0	7	0	4	0	3	10

Tableau 3-10 Nombre de données aberrantes pour les camions

Intersection	ne	ns	nw	es	ew	en	sw	sn	se	wn	we	ws
107002	30	70	11	130	0	6	38	**105**	29	8	2	**124**
107003	3	8	57	5	0	2	22	19	5	12	0	8
107005	13	3	23	**294**	0	19	6	1	**237**	3	0	2
107006	64	40	11	76	0	71	55	52	32	38	0	**122**
107007	61	15	7	14	14	58	6	2	6	11	6	21
107008	49	28	**140**	20	61	97	5	0	53	**170**	90	0
107010	38	2	13	12	15	28	81	0	**105**	1	1	0
107011	50	3	14	2	18	46	64	2	12	42	32	79
107012	12	74	1	**149**	0	15	9	1	14	3	0	17
107013	**353**	61	**131**	**184**	1	**112**	**106**	78	**225**	**137**	1	**142**
107014	20	0	4	24	0	25	1	1	7	2	0	3
107015	16	30	29	37	0	5	28	7	9	23	0	**107**
107016	1	0	16	0	16	2	13	0	1	23	11	12
107017	0	0	0	8	0	3	61	80	72	1	0	73
107018	0	0	0	7	0	0	2	0	23	0	0	4

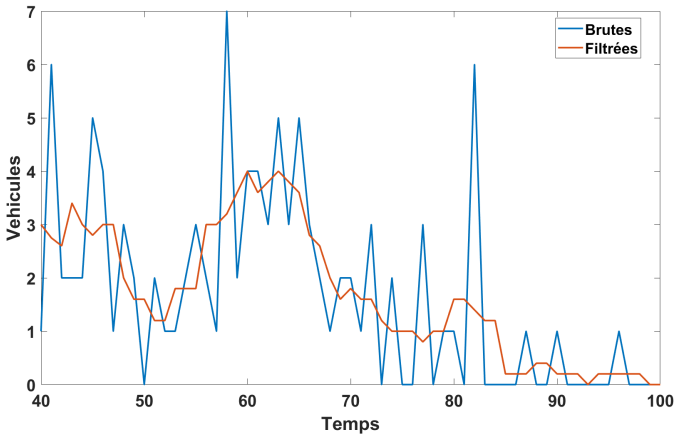


Figure 3.22 Intersection 107002 : véhicules direction Est-Nord

3.4 Conclusion

L'analyse exploratoire des données a permis de dégager plusieurs constats clés sur le comportement des usagers de la route, la distribution des volumes, ainsi que les caractéristiques des conflits observés. Elle met en lumière des tendances significatives et fournit les bases nécessaires à l'élaboration d'indicateurs pertinents pour l'évaluation de la sécurité et de la fluidité aux intersections.

Ces constats serviront de fondement à l'élaboration des méthodes de modélisation et de prédiction qui feront l'objet du chapitre suivant. En effet, le chapitre 4 présentera donc la méthodologie proposée, incluant les choix de modèles, les algorithmes utilisés, et la démarche suivie pour transformer les données brutes en outils d'aide à la décision pour les gestionnaires de la mobilité urbaine.

Chapitre 4 - Méthodes récursives pour la prédiction de la fluidité

Ce chapitre est consacré à l'élaboration de la modélisation du système étudié. L'objectif principal est de construire une représentation formelle et cohérente des phénomènes observés aux intersections urbaines, en lien avec les enjeux de sécurité et de fluidité du trafic. La modélisation permet non seulement de structurer les données collectées, mais aussi de guider le choix des algorithmes d'analyse et de prédiction utilisés dans les étapes ultérieures.

Dans un premier temps, les variables pertinentes sont identifiées, catégorisées et reliées selon une logique causale ou corrélative. Cette étape implique la formulation d'hypothèses de modélisation fondées sur les observations empiriques et les résultats de l'analyse exploratoire des données (chapitre 3).

Sur le plan technique, le chapitre détaille les méthodes statistiques et algorithmiques retenues pour modéliser les dynamiques temporelles et comportementales. À cet effet, les propriétés d'autocorrélation et de corrélation partielle détectées (sections 3.1.3 et 3.1.4) motivent le recours à une modélisation hybride combinant des modèles SARIMA pour capturer les régularités temporelles et le filtre d'estimation adaptatif AEKF pour intégrer les aspects non linéaires et les incertitudes sur les états du système.

Cette approche hybride permet de tirer parti des forces respectives des deux méthodes : les modèles SARIMA sont performants pour décrire les tendances saisonnières et les effets cycliques, tandis que le filtre AEKF est bien adapté pour estimer des états latents dans un contexte dynamique et bruité, tel que celui de la circulation urbaine.

Stratégie de combinaison

Le modèle SARIMA a été utilisé pour effectuer des prévisions sur les données de circulation des véhicules particuliers. Ce modèle permet de capturer à la fois les tendances à court terme et les composantes saisonnières présentes dans les données, ce qui est particulièrement pertinent dans un contexte où la circulation présente des schémas répétitifs quotidiens ou hebdomadaires.

Cependant, les modèles SARIMA présentent certaines limitations, notamment lorsqu'ils sont confrontés à des variations soudaines ou à des événements imprévus. C'est pourquoi une approche hybride combinant SARIMA et un filtre de Kalman étendu adaptatif, *Adaptive Extended Kalman Filtering* (AEKF) a été proposée.

Dans cette combinaison, le modèle SARIMA fournit une estimation de la tendance principale, tandis que l'AEKF ajuste dynamiquement cette estimation en fonction des erreurs d'observation réelles. Cela permet une meilleure réactivité du modèle face aux anomalies ou changements brusques de trafic, tout en conservant les avantages d'un modèle saisonnier robuste.

Cette méthode hybride permet ainsi d'améliorer la précision des prévisions de trafic en exploitant les forces complémentaires des deux approches.

4.1 Prédiction de flux routier par combinaison SARIMA et filtre adaptatif de Kalman

Le filtre de Kalman est un algorithme récursif permettant d'estimer l'état d'un système dynamique en temps réel à partir de mesures bruitées. Il est optimal pour les systèmes linéaires gaussiens.

L'équation d'état et l'équation d'observation sont définies par :

$$x_{k+1} = F_k x_k + B_k u_k + w_k \tag{4.1}$$

$$y_k = H_k x_k + v_k \tag{4.2}$$

où: x_k état du système à l'instant k, y_k observation à l'instant k, u_k le vecteur d'entrée du système, F_k matrice de transition d'état, B_k matrice de contrôle, H_k matrice d'observation et w_k et v_k bruits de processus et de mesure (gaussiens). Le filtre de Kalman étendu est une version adaptée pour les systèmes non linéaires. Il linéarise localement le système à l'aide de la matrice jacobienne.

Les deux équations sont également définies par :

$$x_{k+1} = f(x_k, u_k) + w_k \tag{4.3}$$

$$y_k = h(x_k) + v_k \tag{4.4}$$

$$F_k = \left.\frac{\partial f}{\partial x}\right|_{x=\hat{x}^k} \tag{4.5}$$

$$H_k = \left.\frac{\partial h}{\partial x}\right|_{x=\hat{x}^k} \tag{4.6}$$

.

L'AEKF améliore encore l'EKF en adaptant dynamiquement les matrices de bruit Q (processus) et R (observation) pour compenser les changements inattendus dans l'environnement.

Initialisation : À cette étape, un état initial : $\hat{x}_0$, la matrice de covariance initiale : P_0 et les matrices de bruit initiales : Q_0 et R_0 sont définies.

Prédiction : calcule *a priori* l'état $\hat{x}_k \mid_{k-1}$ et la covariance $P_k \mid_{k-1}$ en appliquant leurs valeurs à $k-1$ à (4.7) et (4.8) respectivement :

$$\hat{x}_k \mid_{k-1} = f(\hat{x}_{k-1}, u_{k-1}) \tag{4.7}$$

$$P_k \mid_{k-1} = F_k P_{k-1} F_k^T + Q_k) \tag{4.8}$$

Mise à jour adaptative : prend en compte la mesure pour calculer les matrices de bruit Q_k et R_k respectivement grâce à (4.9) et (4.10) :

$$Q_k = \alpha Q_{k-1} + (1-\alpha) e_k e_k^T \tag{4.9}$$

$$R_k = \beta R_{k-1} + (1-\beta)(y_k - h(\hat{x}_k \mid k-1))(y_k - h(\hat{x}_k \mid k-1))^T \tag{4.10}$$

Mise à jour : calcule le gain de Kalman K_k avec (4.11) et ajuste les prédictions *a postiori* de l'état $\hat{x}_k$ et de la covariance $P_k = (I - K_k H_k) P_k \mid k-1$ grâce à (4.12) et (4.13) respectivement.

$$K_k = P_k \mid k-1 H_k^T (H_k P_k \mid k-1 H_k^T + R_k)^{-1} \tag{4.11}$$

$$\hat{x}_k = \hat{x}_k \mid k-1 + K_k (y_k - h(\hat{x}_k \mid k-1)) \tag{4.12}$$

$$P_k = (I - K_k H_k) P_k \mid k-1 \tag{4.13}$$

4.2 Prédiction de flux routier par filtre adaptatif linéaire des moindres carrés récursifs

Le filtre adaptatif linéaire des moindres carrés récursifs RLS est une méthode puissante de filtrage adaptatif qui peut être utilisée pour prédire les flux routiers dans un environnement dynamique. Il permet de modéliser la relation entre les entrées (par exemple, des facteurs comme la température, l'humidité, ou le jour de la semaine) et les sorties (flux de circulation, par exemple) en s'adaptant continuellement aux nouvelles données.

Le filtre RLS minimise l'erreur quadratique entre la sortie estimée du modèle et la sortie réelle, en ajustant les poids (ou coefficients) du modèle en temps réel. Contrairement au filtre LMS, qui est basé sur une approximation stochastique, le RLS utilise une méthode exacte pour minimiser cette erreur, ce qui le rend plus rapide et plus précis, bien qu'il soit aussi plus coûteux en termes de calcul.

Initialisation des paramètres :

- État $x(0)$ (initialisation à zéro).
- Poids $w(0)$ (initialisation à zéro).
- Matrice de covariance inverse $P(0)$, souvent initialisée à une grande valeur pour refléter une incertitude initiale élevée.

Pour chaque nouvel échantillon k :

1. Calculer l'erreur avec (4.14):

$$e_k = y_k - x_k^T w(k-1) \tag{4.14}$$

où y_k est la sortie réelle et x_k est le vecteur d'entrée.

2. Mettre à jour la matrice de covariance inverse $P(k)$ en appliquant (4.15):

$$P(k) = P(k-1) - \frac{P(k-1)x_k x_k^T P(k-1)}{1 + x_k^T P(k-1)x_k} \tag{4.15}$$

3. Mettre à jour les poids $w(k)$ en appliquant (4.16) :

$$w(k) = w(k-1) + \frac{e_k P(k-1)x_k}{1 + x_k^T P(k-1)x_k} \tag{4.16}$$

À chaque nouvel échantillon, il faut répéter ces étapes.

4.3 Métriques utilisées pour l'évaluation des modèles

Il est important de se définir des métriques pour quantifier les résultats. À cet effet, trois métriques sont choisies : le *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE) et le R^2.

RMSE ou la racine de l'erreur quadratique moyenne

Cette métrique est définie par (4.17) :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.17}$$

où y_i sont les valeurs réelles et $\hat{y}_i$ sont les valeurs prédites par le modèle, et n est le nombre d'observations.

L'erreur RMSE mesure l'écart moyen entre les valeurs prédites et les valeurs réelles. Il est particulièrement sensible aux grandes erreurs en raison de la présence du carré dans la formule. Cela signifie que les grandes erreurs auront un impact plus important sur cette métrique. Cette caractéristique est utile si les erreurs importantes doivent être

particulièrement pénalisées dans votre contexte. Une erreur RMSE plus faible indique une meilleure précision globale du modèle.

MAE ou Erreur Absolue Moyenne

L'équation (4.18) définit la métrique :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.18}$$

où y_i et $\hat{y}_i$ sont respectivement les valeurs réelles et les valeurs prédites.

L'erreur MAE mesure l'écart moyen absolu entre les valeurs réelles et prédites, sans donner de poids supplémentaire aux grandes erreurs, contrairement au RMSE. Cette métrique est souvent utilisée lorsqu'une évaluation globale de la performance du modèle est souhaitée, pénalisant toutes les erreurs de manière égale, indépendamment de leur taille. Un MAE plus faible signifie une meilleure performance globale du modèle, sans accorder une attention particulière aux erreurs importantes.

R^2 *(Coefficient de Détermination)*

Le coefficient de détermination R^2 se définit ainsi par (4.19)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.19}$$

où $\bar{y}$ est la moyenne des valeurs réelles y_i.

Le coefficient de détermination R^2 mesure la proportion de la variance des données réelles qui est expliquée par le modèle. Il varie de 0 à 1, où un R^2 de 1 signifie que le modèle explique parfaitement la variance des données, et un R^2 de 0 signifie qu'il n'explique aucune

des variations. Un R^2 plus élevé indique que le modèle est plus efficace pour expliquer la variabilité des données, ce qui est crucial dans l'évaluation de la performance, notamment dans les séries temporelles.

Justifications des métriques

Le RMSE est particulièrement utile si vous souhaitez pénaliser les grandes erreurs, ce qui peut être important dans des systèmes où des erreurs significatives peuvent avoir un impact important. Le MAE offre une vue plus "équilibrée" de l'erreur moyenne, sans accorder un poids supplémentaire aux erreurs plus grandes, ce qui est utile pour évaluer la performance globale du modèle de manière uniforme. Le R^2 permet de quantifier dans quelle mesure le modèle explique la variabilité des données, ce qui est particulièrement important dans les tâches de régression où la compréhension de la qualité du modèle par rapport à la variabilité totale est essentielle.

En résumé, ces trois métriques offrent une vision complète de la performance du modèle: le RMSE et le MAE pour les erreurs directes, et le R^2 pour l'ajustement global du modèle aux données.

4.4 Résultats de la prédiction de la fluidité

4.4.1 Prédictions du modèle SARIMA

Le modèle SARIMA a été utilisé pour effectuer des prévisions sur les données de circulation des véhicules particuliers (intersection **107002**). Ce modèle permet de capturer à la fois les tendances à court terme et les composantes saisonnières présentes dans les données, ce qui est particulièrement pertinent dans un contexte où la circulation présente des schémas répétitifs quotidiens ou hebdomadaires [21].

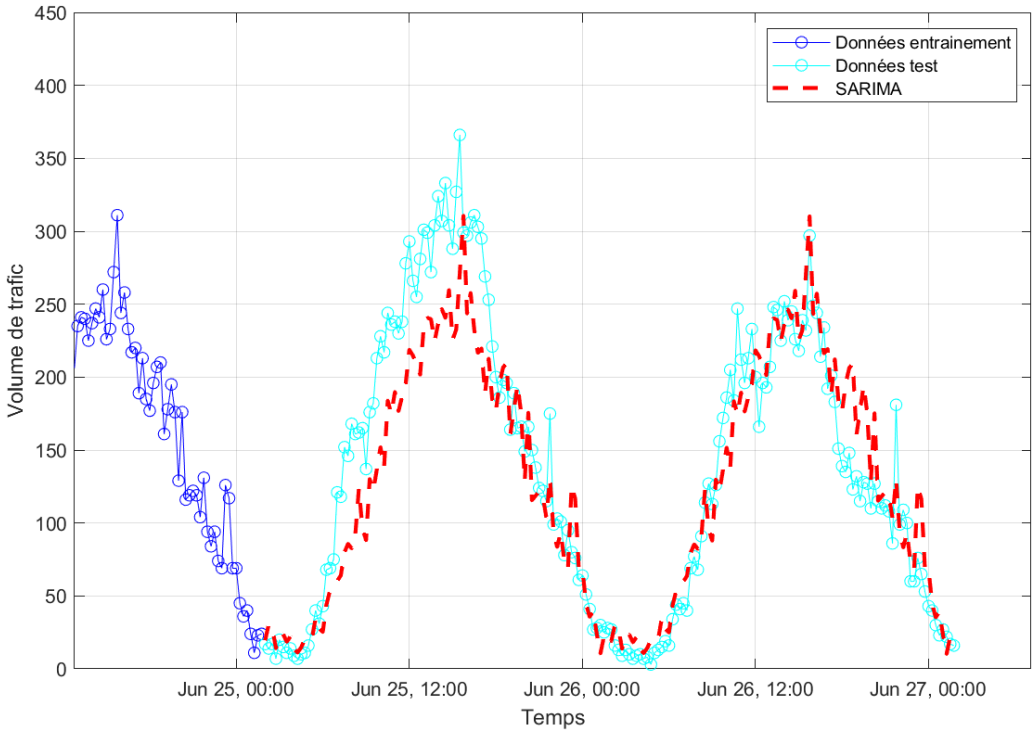


Figure 4.1 Prédiction du modèle SARIMA

Un modèle simple mais robuste, souvent utilisé en pratique : SARIMA$(1,0,1) \times (1,1,1,96)$. $d = 0$ signifie qu'il n'y a pas de différenciation nécessaire si la série est déjà stationnaire. $D = 1$ pour une différenciation saisonnière sur 96 points (pour gérer les pics quotidiens). Des valeurs de $p = 1$, $q = 1$ permettent de modéliser les effets d'inertie à court terme. Quant aux valeurs de $P = 1$, $Q = 1$, elles permettent de modéliser les effets d'inertie saisonniers. La Figure 4.1 montre une bonne détection de la saisonnalité.

Cependant, les modèles SARIMA ont certaines limitations, notamment lorsqu'ils sont confrontés à des variations soudaines ou à des événements imprévus. C'est pourquoi une approche hybride combinant SARIMA et un filtre de Kalman étendu adaptatif (AEKF) a été proposée.

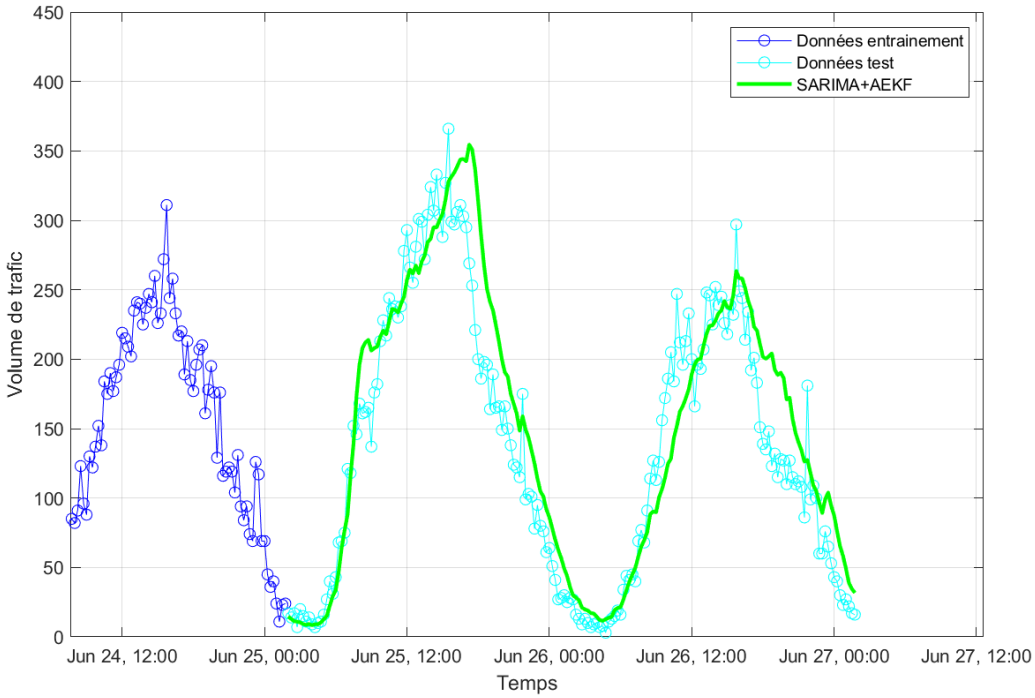


Figure 4.2 Prédiction du modèle hybride SARIMA et AEKF

4.4.2 Prédictions du modèle combiné SARIMA et AEKF

Dans cette combinaison, le modèle SARIMA fournit une estimation de la tendance principale, tandis que l'AEKF ajuste dynamiquement cette estimation en fonction des erreurs d'observation réelles. Cela permet une meilleure réactivité du modèle face aux anomalies ou changements brusques de trafic, tout en conservant les avantages d'un modèle saisonnier robuste.

Cette méthode hybride permet ainsi d'améliorer la précision des prévisions de trafic en exploitant les forces complémentaires des deux approches. L'approche hybride SARIMA + AEKF offre une meilleure précision de prévision en combinant la robustesse des modèles saisonniers classiques et l'adaptabilité des filtres bayésiens non linéaires. Elle est particulièrement efficace dans les scénarios présentant des ruptures ou des anomalies dans le trafic.

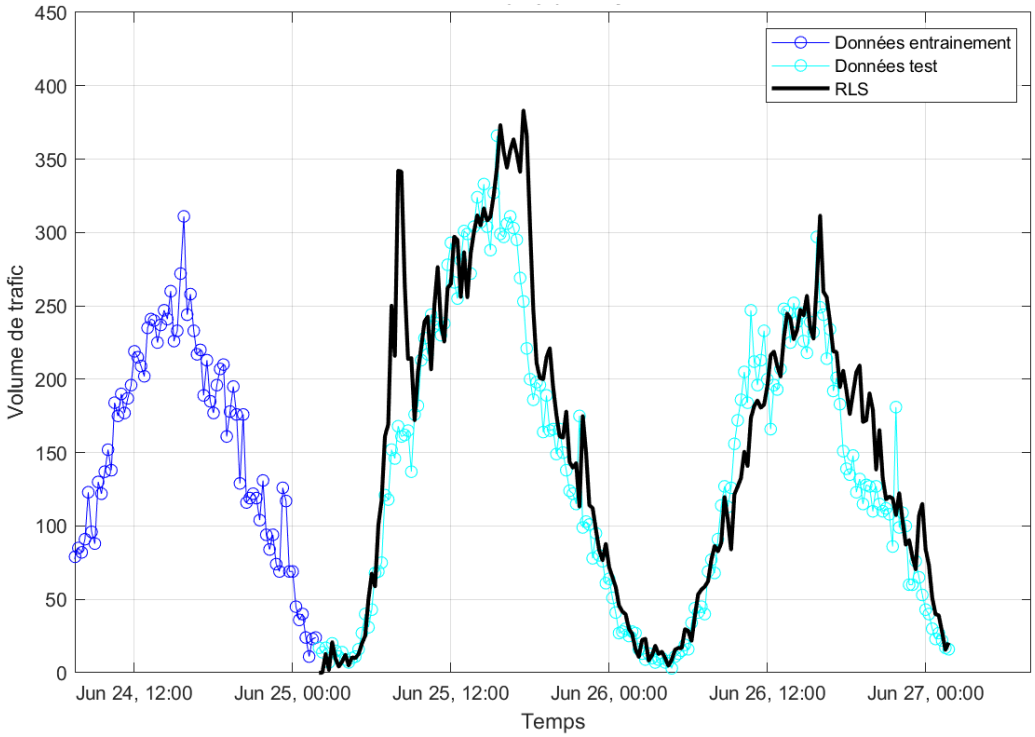


Figure 4.3 Prédiction du modèle RLS

4.4.3 Prédictions par filtre linéaires RLS

À la Figure 4.3, la ligne noire représente les prédictions du modèle RLS comme indiqué. Elle suit de près les données de test, ce qui indique une bonne capacité du modèle à capturer les tendances du trafic routier. La méthode RLS semble efficace pour prédire les volumes de trafic, même sur des périodes prolongées. Elle est rapide et adaptée aux environnements dynamiques où les conditions de trafic changent fréquemment. La précision peut diminuer si les données d'entraînement ne reflètent pas bien les conditions réelles, ce qui explique la pertinence d'utiliser une élimination des aberrances.

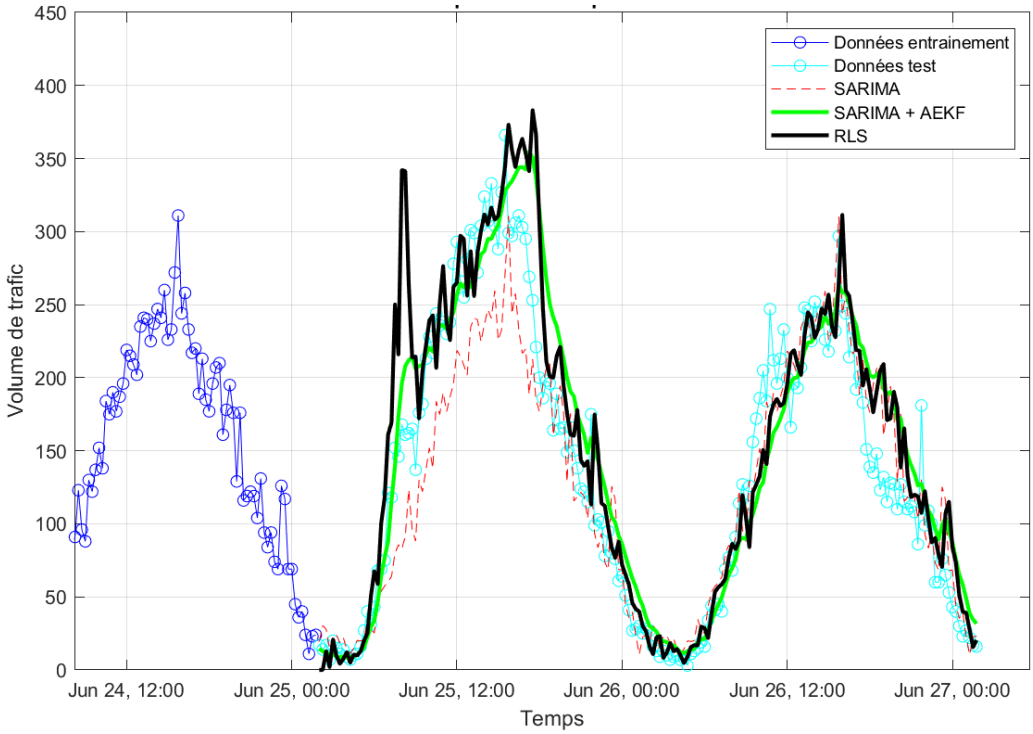


Figure 4.4 Comparaison des prédictions

Tableau 4-1 Résultats des tests de prédictions de flux de véhicules

	Test		
	RMSE	*MAE*	R^2
SARIMA	*38.04*	*28.19*	*0.84*
SARIMA+AEKF	*36.38*	*27.52*	*0.85*
RLS	*40.71*	*27.84*	*0.82*

4.4.4 Analyse de la précision des résultats et discussion

Les résultats obtenus à partir des modèles SARIMA, SARIMA-AEKF et AEKF-RLS montrent des différences significatives en matière de performance prédictive. L'évaluation a été faite sur plusieurs tronçons et périodes de la journée (matin, après-midi, soir) à l'aide des métriques usuelles : RMSE, MAE et R^2.

Les performances du modèle hybride SARIMA-AEKF se démarquent de manière constante pour les prédictions à court terme (15 à 30 minutes), notamment grâce à sa capacité à capter à la fois les composantes linéaires saisonnières (SARIMA) et les dynamiques adaptatives (AEKF). Les résultats montrent un R^2 supérieur à 0.85 dans la majorité des cas, contre une moyenne de 0.72 pour le SARIMA seul et 0.78 pour l'AEKF-RLS. Cela confirme la pertinence d'une approche hybride dans le contexte de données de trafic fluctuantes et bruitées.

Du point de vue opérationnel, une telle précision permettrait aux autorités municipales de mieux anticiper les pics de trafic, d'optimiser les feux de signalisation, ou encore de planifier la présence de personnel pour les interventions préventives. De plus, les modèles sont robustes aux ruptures locales, ce qui est essentiel en milieu urbain sujet aux événements ponctuels (travaux, incidents).

Les limites rencontrées concernent notamment la sensibilité aux valeurs aberrantes, présentes dans certaines périodes de fin de journée, où une congestion inattendue provoque des écarts importants. Une piste d'amélioration serait l'intégration de variables exogènes (météo, événements spéciaux), ce qui sera exploré dans les perspectives futures.

Ces observations confortent l'idée que l'intelligence artificielle, même dans des modèles simples ou semi-paramétriques, apporte une valeur ajoutée immédiate à la gestion proactive des infrastructures urbaines.

4.5 Conclusion

Dans ce chapitre, nous avons proposé et évalué une méthode hybride de prédiction du volume de trafic à court terme, combinant des modèles statistiques et adaptatifs. Les

résultats obtenus à partir des données réelles de la Ville de Trois-Rivières démontrent l'efficacité de cette approche, en particulier dans un contexte urbain dynamique.

La capacité du modèle à fournir des prédictions fiables constitue un levier précieux pour une gestion plus intelligente et efficiente des ressources urbaines. L'intégration de telles solutions dans les systèmes de transport intelligents constitue un pas de plus vers une ville plus fluide, plus sécuritaire et plus durable.

Dans le chapitre suivant, l'analyse se concentrera sur les enjeux liés à la sécurité routière, à travers l'utilisation de mesures indirectes, telles que le TPA, pour identifier et prédire les situations à risque dans les intersections urbaines. Nous verrons comment des techniques similaires d'intelligence artificielle peuvent aussi être appliquées pour évaluer le niveau de risque d'une intersection à partir des données de conflits usagers.

Chapitre 5 - Méthodes par apprentissage automatique sur la sécurité

L'évaluation des risques aux intersections urbaines est un enjeu clé pour la gestion du trafic et la sécurité routière. Cette section propose une approche basée sur l'apprentissage automatique pour classer en temps réel le niveau de risque des intersections. L'étude repose sur trois étapes principales : 1) La détection des anomalies avec (*Density Based Spatial Clustering Applications with Noises* DBSCAN) pour identifier les événements de trafic atypiques. 2) L'estimation du risque avec une méthode statistique d'une distribution de Pareto généralisée (*Generalized Pareto Distribution* - GPD) appliquée aux dépassements de seuil. 3) Enfin, la classification du niveau de risque en utilisant des modèles d'apprentissage automatique, incluant un réseau de neurones optimisé (*Optimized Neural Network* - ONN), séparateur à vaste marge optimisé (*Optimized Support Vector Machine* - OSVM), une régression logistique efficace (ELR) et un modèle naïf bayésien (GNB).

Les intersections urbaines sont des points critiques en matière de sécurité routière. La classification des risques en temps réel permettrait aux gestionnaires de la circulation d'anticiper les zones dangereuses et d'optimiser la prévention des accidents.

Les approches traditionnelles basées sur l'historique des accidents nécessitent de longues périodes d'observation. À l'inverse, les indicateurs de sécurité substitutifs, tels que le TPA et la vitesse des usagers, offrent une évaluation plus dynamique des risques. Ce travail propose un modèle hybride combinant l'analyse des anomalies et l'apprentissage automatique pour une classification efficace des risques aux intersections.

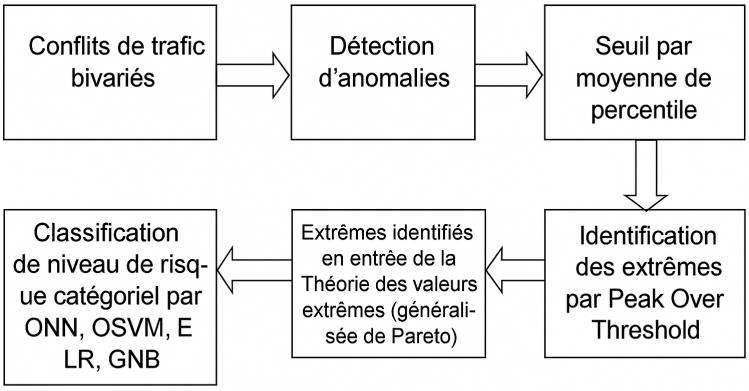


Figure 5.1 Methodologie proposée pour l'indice de sécurité

5.1 Détection d'anomalie

L'algorithme DBSCAN répond aux limites des algorithmes classiques de regroupement, tels que le K-means, qui sont inadaptés aux données présentant des formes complexes, comme indiqué dans [22].

Dans cette étude, l'algorithme DBSCAN est utilisé pour regrouper les points de conflit en deux catégories : événements normaux et événements anormaux. Grâce à son approche basée sur la densité, DBSCAN permet d'identifier efficacement les événements rares et à haut risque.

Après analyse de la structure de nos données, la notion intuitive de regroupements (*clusters*) et de « bruit » (données aberrantes) décrite dans [22] s'applique naturellement. En effet, les données présentent une forte densité de points autour de certaines valeurs de TPA et de vitesse, correspondant probablement à des comportements normaux ou habituels lors des interactions entre usagers de la route. En revanche, les conflits graves, susceptibles d'entraîner des accidents mortels, sont plus rares et se distinguent par leur faible densité.

Comme l'illustre la distribution conjointe du TPA et de la vitesse dans la figure 5.2, ces points de données représentant des conflits dangereux sont identifiés comme du bruit par l'algorithme DBSCAN, conformément à sa définition et à la faible fréquence attendue de ces événements.

Plutôt que d'utiliser une méthode classique d'analyse de stabilité des paramètres pour définir un seuil, nous avons adopté une approche alternative. Le ratio de points bruités par rapport au nombre moyen de points regroupés est fixé comme étant le pourcentage d'un percentile spécifique. Cette méthode nous permet de définir efficacement le seuil optimal. Les paramètres de DBSCAN, à savoir : λ (nombre minimal de points pour former un cluster) et ε (distance maximale entre deux points pour qu'ils soient considérés comme voisins). Ils sont déterminés empiriquement par une recherche par grille. Cette démarche vise à équilibrer la sensibilité de la détection des anomalies et la réduction des faux positifs.

Cependant, cette approche présente une limitation : les conditions de trafic varient d'une intersection à l'autre. Par conséquent, l'application des mêmes paramètres DBSCAN à toutes les intersections peut parfois limiter les performances du modèle.

5.2 Méthodes statistiques : Distribution de Pareto généralisée

Conformément à l'approche proposée par [1], cette étude s'appuie sur les formulations et les fondements théoriques présentés dans [23].

L'indice de risque est défini par (5.1) :

$$R_i x = Pr(X_i > \mu) = 1 - G(X_i) \tag{5.1}$$

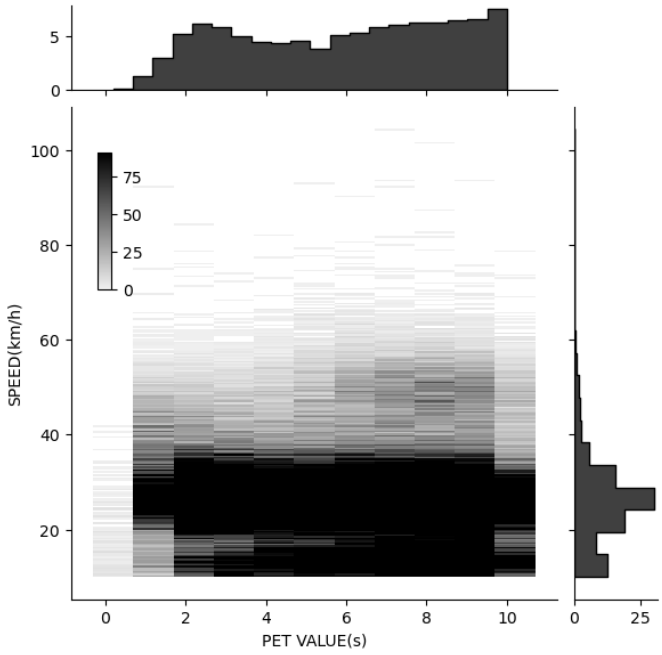


Figure 5.2 Distribution jointe marginale des histogrammes de TPA et vitesse pour les données de l'intersection **107002**

où $G(X_i)$ est la fonction de répartition de la GPD, μ représente le seuil, et X_i est la variable aléatoire.

Ensuite, l'indice de risque (5.1) est appliquée aux données de dépassement du TPA et de la vitesse pour chaque bloc horaire i, permettant ainsi de calculer séparément $R_i p$ et $R_i s$.

L'indice de risque global R_i est alors obtenu en prenant la moyenne de ces deux valeurs, tel que donné par (5.2):

$$R_i = \frac{R_i p + R_i s}{2} \tag{5.2}$$

Enfin, une catégorie de sécurité théorique est définie en fonction de la valeur de R_i calculée.

Les dépassements des valeurs seuils du TPA négatif et de la vitesse sont modélisés à l'aide d'une distribution de Pareto généralisée (GPD). Soit X_1, X_2,..., X_n une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), ayant pour fonction de répartition marginale F. Il est naturel de considérer que certaines des valeurs X_i correspondent à des événements extrêmes, c'est-à-dire qu'elles dépassent un seuil élevé μ.

En désignant une variable arbitraire de la séquence X_i par X, on obtient (5.3) qui est l'équation de probabilité conditionnelle suivante, qui décrit le comportement stochastique des événements extrêmes :

$$Pr\{X > \mu + y \mid X > \mu\} = \frac{1 - F(\mu + y)}{1 - F(\mu)}, \quad y > 0 \tag{5.3}$$

Cette expression représente la probabilité qu'une valeur X dépasse un seuil $\mu + y$, sachant qu'elle a déjà dépassé μ. D'après [23], la fonction GPD associée à (5.1) est donnée par (5.4) :

$$G(x) = \exp\left(-\left[1 + \xi \frac{x - \mu}{\sigma}\right]^{-\frac{1}{\xi}}\right) \tag{5.4}$$

où les paramètres de la distribution sont : le paramètre de forme (ξ, $\xi \neq 0$), le paramètre d'échelle σ et le seuil μ. L'équation (5.4) constitue une caractérisation asymptotique du modèle décrit par l'équation (5.3).

5.3 Apprentissage supervisé et classification du niveau de risque

Toutes les données précédemment collectées concernant le TPA et la vitesse sont disponibles, ce qui facilite le calcul de l'indice de risque conformément à l'approche décrite dans 5.2.

Pour la classification, les caractéristiques des conflits utilisées en entrée sont le TPA et la vitesse. Quatre modèles d'apprentissage automatique sont employés : un réseau de neurones artificiel (*Artificial Neural Network*), une machine à vecteurs de support (*SVM - Support Vector Machine*), une régression logistique efficace (*ELR - Efficient Linear Regression*) et une gaussienne bayesienne naive (*GNB - Gaussian Naïve Bayes*). Parmi ces modèles, le réseau de neurones et la machine à vecteurs de support ont été optimisés pour améliorer leurs performances.

Les paramètres optimisés pour le réseau de neurones sont décrits comme suit. Le nombre de couches entièrement connectées dans l'intervalle 1 et 3. La fonction d'activation parmi ReLU, tanh, sigmoïde ou aucune. La force de régularisation comprise entre 3.31 10^{-10} et 3.31. La taille des première et deuxième couches prise entre 1 et 300. Et une application ou non de standardisation des données. Pour les paramètres optimisés du modèle SVM, nous avons la fonction noyau choisie parmi une gaussienne, une linéaire, une quadratique ou une cubique. Le niveau de contrainte varie entre 0.001 et 1000. L'approche multiclasses One-vs-All ou One-vs-One. Une application ou non d'une standardisation des données.

Définition des catégories de sécurité

La classification du niveau de sécurité est définie par une catégorie verte de risque faible, une catégorie jaune de risque modéré et une catégorie rouge de risque élevé respectivement si la valeur $R_i \leq 0.35$, $0.35 < R_i \leq 0.65$, $R_i > 0.65$. Chaque point de

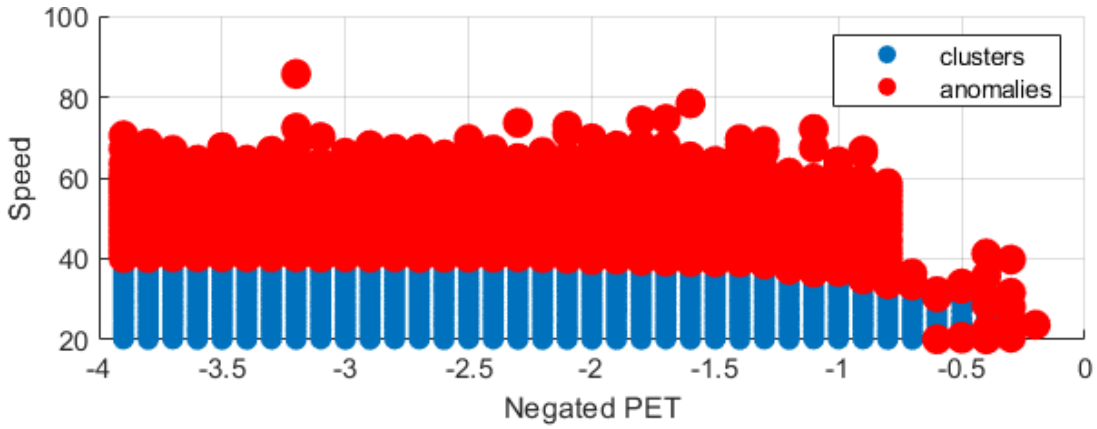


Figure 5.3 Détection d'anomalie par DBSCAN

données de conflit est traité individuellement comme entrée du modèle de classification. Les seuils de classification ont été déterminés empiriquement via une analyse de sensibilité afin d'optimiser la précision du modèle tout en garantissant une différenciation significative entre les niveaux de risque.

La section suivante détaille la mise en œuvre opérationnelle de cette modélisation à travers des simulations et des tests en discutant des résultats.

5.4 Résultats d'apprentissage automatique

5.4.1 Détection d'anomalie

L'objectif principal de cette étude est de prédire avec précision le niveau de sécurité (catégorie) des intersections de la ville de Trois-Rivières. Toutes les simulations et implémentations ont été réalisées à l'aide des outils MATLAB. Pour la détection des anomalies, les paramètres du DBSCAN sont définis par $\varepsilon = 0.45$ et $\gamma = 220$. Comme illustré dans la Figure 5.3, les points de données anormaux identifiés correspondent à ceux de la distribution conjointe présentée dans la Figure 5.2.

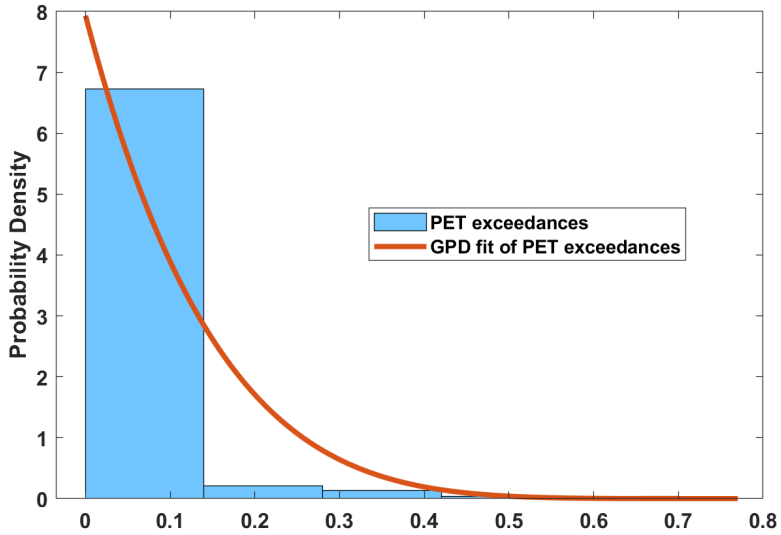


Figure 5.4 Ajustements GPD des excès de TPA

Tableau 5-1 Résultats des paramètres d'ajustements GPD

Ajustement TPA négatif			**Ajustement vitesse**		
Forme:ξ	*Échelle:*σ	*Seuil:*μ	*Forme:*ξ	*Échelle:*σ	*Seuil:*μ
-0.16	0.12	-0.9	0.19	3.04	56.2

5.4.2 Niveau de risque de sécurité

Ajustement de la distribution par GPD

L'utilisation de la GPD est motivée par la décroissance exponentielle observée dans la queue de la distribution des données [24], en particulier pour le TPA négatif. Les paramètres de la GPD sont estimés via la méthode du maximum de vraisemblance (MLE), avec un intervalle de confiance de 95%, garantissant ainsi un ajustement optimal aux données.

Aux figures 5.4 et 5.5, l'histogramme bleu représente la densité de probabilité des dépassements de la TPA, capturant les valeurs sous le seuil (dans le cas des valeurs négatives). Les barres indiquent la fréquence de ces dépassements, normalisée pour refléter

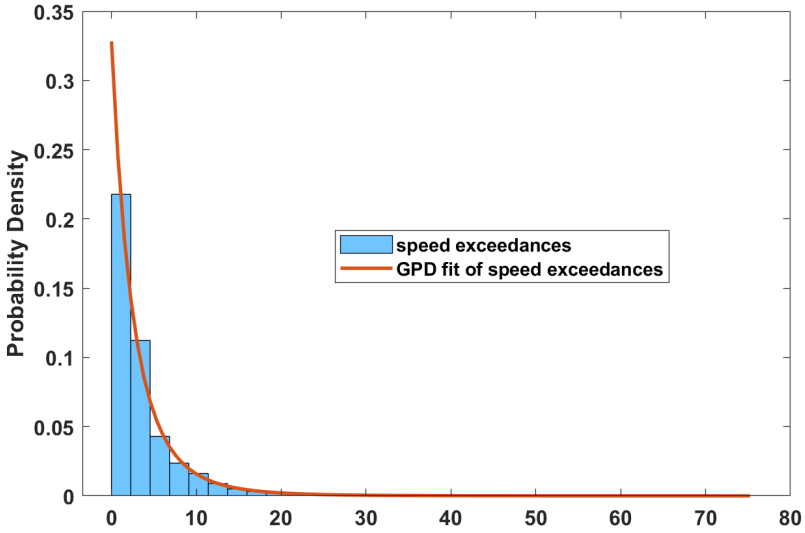


Figure 5.5 Ajustements GPD des excès de vitesse

une densité de probabilité. La ligne rouge montre l'ajustement du GPD appliqué aux données de dépassement. L'augmentation abrupte près de zéro suggère que la distribution est fortement concentrée autour des faibles valeurs de dépassement, une caractéristique couramment observée dans les distributions ayant des queues de type Pareto.

Le fort alignement entre l'histogramme et la courbe rouge démontre un bon ajustement, validant ainsi l'utilisation du GPD pour modéliser les données de dépassement de la TPA. Cela confirme l'adéquation du GPD pour capturer les valeurs extrêmes de la TPA. La forme de la courbe suggère que le paramètre de forme ξ est probablement proche de zéro, ce qui est caractéristique des distributions à queues décroissantes polynomiales, couramment observées dans l'analyse des valeurs extrêmes. Cependant, de légères divergences entre le GPD ajusté et l'histogramme dans la figure 5.4 apparaissent en raison de la faible densité de données dans les régions de valeurs extrêmes.

La tableau 5-1 présente un résumé des résultats d'ajustement pour la TPA et la vitesse. Le paramètre de forme (ξ = -0.16) indique une queue bornée, suggérant une gamme limitée de valeurs extrêmes. Cependant, de légères divergences entre le GPD ajusté et l'histogramme dans la figure 5.4 apparaissent en raison de la rareté des données dans les régions de valeurs extrêmes.

Dans la figure 5.5, l'histogramme bleu représente la densité de probabilité des dépassements de vitesse. Les barres indiquent la fréquence de ces dépassements, normalisée pour refléter une densité de probabilité. La ligne rouge montre l'ajustement du GPD appliqué aux données de dépassement de vitesse. Le paramètre de forme positif (ξ = 0.19) indique une distribution à queue lourde, suggérant une probabilité plus élevée de valeurs extrêmes.

Classification

Pour la classification, 75% des données sont allouées à l'entraînement et à la validation, comprenant 17461 classes *vert*, 6976 classes *jaunes* et 808 classes *rouge*, une validation croisée est appliquée pour atténuer le surapprentissage. Les 25% restants sont réservés pour les tests, comprenant 2071 classes *vert*, 4701 classes *jaunes* et 1641 classes *rouge*. Comme indiqué dans la tableau 5-2, la précision globale est utilisée comme métrique clé de performance, où une précision plus élevée indique une meilleure performance de classification. De plus, le coût total représente le nombre d'instances mal classées (plus bas est meilleur), tandis que le taux d'erreur sert d'indicateur de la tendance du modèle à faire des erreurs de classification. Les métriques sont obtenues par les équations (5.5) à (5.7).

$$Precis. = \frac{N_{TPR}}{N_{TOTAL}} \tag{5.5}$$

$$CoutTot. = N_{FNR} \tag{5.6}$$

$$TauxErr = 1 - Precis. \tag{5.7}$$

où N_{TOTAL}, N_{TPR} et N_{FNR} représentent respectivement le nombre total d'observations, le nombre de prédictions correctes et le nombre de prédictions incorrectes.

Des informations supplémentaires peuvent être obtenues à partir de métriques telles que le taux de vrais positifs (*TPR - True Positive Rate*), qui quantifie la proportion d'instances correctement prédites dans chaque classe, et le taux de faux négatifs (*FNR - False Negative Rate*), qui mesure la proportion d'instances négatives mal classées. Les matrices de confusion présentées dans les tableaux 5-3 et 5-4 détaillent la performance de classification pour les ensembles de validation et de test à travers différentes méthodes.

La classification comparative est effectuée en utilisant quatre modèles : ONN, OSVM, ELR et GNB. Les métriques d'évaluation incluent l'exactitude (pour les ensembles de validation et de test), le TPR, le FNR et les matrices de confusion pour évaluer la performance des modèles.

Les résultats montrent que ONN est le modèle le plus performant avec une précision de 79.37% en test, permettant ainsi une classification fiable du niveau de dangerosité des intersections.

La majorité des points de données de TPA et de vitesse appartiennent à la catégorie *yellow*, ce qui indique une haute précision de classification pour cette classe, comme le montrent les résultats des matrices de confusion dans les tableaux 5-3 et 5-4. Afin

Tableau 5-2 Résultats des tests et validation des méthodes. N_{TOTAL} de 33658 observations, 22720 observations pour l'entrainement, 2525 pour la validation et 8413 observations pour le test

	Validation			**Test**		
	Précis.	**Coût Tot.**	**Taux Err.**	**Précis.**	**Coût Tot.**	**Taux Err.**
ONN	*78.95%*	*6340*	*21.05%*	*79.37%*	*5084*	*20.63%*
OSVM	*77.69%*	*6720*	*22.31%*	*78.88%*	*5205*	*21.12%*
ELR	*76.82%*	*6983*	*23.18%*	*78.14%*	*5387*	*21.86%*
GNB	*75.82%*	*7285*	*24.18%*	*77.72%*	*5491*	*22.28%*

Tableau 5-3 Résultats de la matrice de confusion de la validation. Nombre d'observation égal à 2525

ONN		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*2225**(55.5%)***	*1691(42.2%)*	*91(2.3%)*	***44.5%***
	Jaune	*685(3.8%)*	*15759(88.2%)*	*1426(8%)*	*11.8%*
	Rouge	*156(1.9%)*	*2291(27.8%)*	*5799**(70.3%)***	***29.7%***
OSVM		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*1787(44.6%)*	*2155(53.8%)*	*65(1.6%)*	*55.4%*
	Jaune	*360(2%)*	*16546(88.2%)*	*964(5.4%)*	*7.4%*
	Rouge	*142(1.7%)*	*3034(36.8%)*	*5070(61.5%)*	*38.5%*
ELR		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*1966(49.1%)*	*2014(50.3%)*	*27(0.7%)*	*50.9%*
	Jaune	*481(2.7%)*	*16459(92.1%)*	*930(5.2%)*	*7.9%*
	Rouge	*193(2.3%)*	*3338(40.5%)*	*4715(57.2%)*	*42.8%*
GNB		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*1835(45.8%)*	*2137(53.3%)*	*35(0.9%)*	*54.2%*
	Jaune	*492(2.8%)*	***16588(92.8%)***	*790(4.4%)*	***7.2%***
	Rouge	*187(2.3%)*	*3644(44.2%)*	*4415(53.5%)*	*46.5%*

de résoudre ce déséquilibre, l'ensemble de données devrait être ajusté pour assurer une distribution plus uniforme entre toutes les classes.

Tableau 5-4 Résultats de la matrice confusion des tests. Nombre d'observation de test égal à 8413

ONN		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*2945(**71.8%**)*	*1072(26.1%)*	*83(2%)*	***28.2%***
	Jaune	*980(7%)*	*12229(86.8%)*	*886(6.3%)*	*13.2%*
	Rouge	*204(3.2%)*	*1859(28.8%)*	*4388(**68%**)*	***32%***
OSVM		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*2444(59.6%)*	*1605(39.1%)*	*51(1.2%)*	*40.4%*
	Jaune	*501(3.6%)*	*13051(92.6%)*	*543(3.9%)*	*7.4%*
	Rouge	*145(2.2%)*	*2360(36.6%)*	*3946(61.2%)*	*38.8%*
ELR		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*2643(64.5%)*	*1438(35.1%)*	*19(0.5%)*	*35.5%*
	Jaune	*648(4.6%)*	*12994(92.2%)*	*453(3.2%)*	*7.8%*
	Rouge	*199(3.1%)*	*2630(40.8%)*	*3622(56.1%)*	*43.9%*
GNB		**Prédictions**			
		Vert	**Jaune**	**Rouge**	**FNR**
Réelles	**Vert**	*2321(56.6%)*	*1740(42.4%)*	*39(1%)*	*43.4%*
	Jaune	*435(3.1%)*	*13226(**93.8%**)*	*434(3.1%)*	***6.2%***
	Rouge	*174(2.7%)*	*2669(41.4%)*	*3608(55.9%)*	*44.1%*

Parmi les modèles testés, ONN a obtenu la meilleure précision, atteignant 78.95% pour la validation et 79.37% pour les tests. Il est suivi par OSVM avec 77.69% et 78.88%, respectivement, puis ELR avec 76.82% et 78.14%, et enfin GNB avec 75.82% et 77.72%.

Les paramètres optimaux pour le modèle ONN incluent : 2 couches entièrement connectées, une fonction d'activation tanh, une force de régularisation de $4.10e^{-10}$, une taille de première couche de 298, une taille de seconde couche de 3 et des données standardisées. Pour OSVM, les paramètres optimisés sont : une fonction de noyau quadratique, un niveau de contrainte de 39.76, un codage multiclass One-vs-One et des données d'entrée standardisées.

L'ONN surpasse les autres modèles grâce à sa capacité à capturer des relations non linéaires, à extraire des caractéristiques hiérarchiques et à optimiser les hyperparamètres pour une meilleure adaptabilité. Sa robustesse dans la classification des scénarios à haut risque (rouge) et à bas risque (vert) en fait le choix le plus fiable pour l'évaluation des risques en temps réel.

Certains pourraient argumenter que le GPD seul est suffisant pour classifier les conflits. Cependant, ses limitations doivent être prises en compte, notamment sa dépendance à des blocs de données complets sur une période donnée avant que la classification puisse être effectuée. Bien que les méthodes de classification dans cette étude utilisent le TPA instantané et la vitesse comme prédicteurs (entrée bivariée), elles fonctionnent sur des points de données individuels plutôt que sur des groupes agrégés, permettant ainsi un traitement en temps réel. De plus, leur précision est toujours évaluée en fonction du cadre GPD (classification théorique du R_i calculé).

5.5 Conclusion

Ce chapitre a introduit une approche basée sur l'apprentissage automatique pour l'évaluation en temps réel des risques aux intersections, exploitant la détection d'anomalies par DBSCAN et un modèle de risque GPD afin de classifier les niveaux de sécurité des intersections. Contrairement aux méthodes traditionnelles fondées sur des données historiques d'accidents, notre approche évalue de manière dynamique des mesures de sécurité substitutives, telles que le PET et les anomalies de vitesse, afin d'améliorer la gestion du trafic et la prévention des accidents. Une des limites de notre approche réside dans l'utilisation de paramètres fixes de DBSCAN pour toutes les intersections, ce qui pourrait réduire son efficacité à détecter des anomalies dans des conditions de trafic variables. Pour y remédier, nous suggérons un modèle unifié de détection d'anomalies qui

s'adapte dynamiquement à différents profils d'intersection en utilisant un seuil généralisé bi-variable. De plus, pour améliorer la détection d'événements rares, DBSCAN pourrait être combiné à une méthode d'apprentissage automatique adaptative qui optimise les paramètres de regroupement en temps réel. Le modèle GPD permet de représenter efficacement la distribution globale des dépassements de PET et de vitesse, fournissant ainsi des informations précieuses pour l'évaluation des risques. Toutefois, ses performances peuvent être moins fiables lorsqu'il est appliqué à une seule intersection. En ce qui concerne les performances de classification, ONN s'est révélé supérieur aux autres modèles, en particulier pour identifier les scénarios à haut risque (rouge) et à faible risque (vert). Le modèle GNB a montré une précision correcte pour la classification des risques modérés (jaune), mais l'ONN est resté le plus efficace dans l'ensemble, atteignant la meilleure précision globale. Des améliorations supplémentaires, comme des techniques d'équilibrage des données et d'ingénierie des caractéristiques, pourraient renforcer la robustesse de la classification et sa généralisation aux conditions réelles. Ces résultats soulignent le potentiel de l'apprentissage automatique dans l'analyse de la sécurité routière et offrent un cadre précieux pour l'intégration de l'évaluation des risques en temps réel dans les systèmes de gestion du trafic urbain. ONN est le meilleur des quatre pour classer le vert et le rouge ; GNB est raisonnable pour classer le jaune, mais ONN reste le meilleur selon la précision globale.

Chapitre 6 - Conclusion

Dans le cadre de ce mémoire, nous avons analysé et comparé trois approches prédictives appliquées au flux routier : SARIMA, SARIMA + AEKF et RLS. Cette étude a permis de mettre en lumière les avantages, les limites, et les performances respectives de ces modèles.

Les résultats obtenus ont démontré que le modèle RLS, grâce à sa flexibilité et sa capacité d'adaptation en temps réel, est particulièrement performant dans des environnements dynamiques. Le modèle SARIMA + AEKF, bien que demandant une complexité algorithmique plus importante, offre une robustesse accrue et des prédictions précises dans des situations où les variations sont modérées. En revanche, le modèle SARIMA classique, bien qu'efficace pour identifier des tendances globales, a montré ses limites face à des conditions de trafic évolutives.

Ces résultats ouvrent la voie à plusieurs perspectives pour de futures recherches. Une intégration hybride des méthodes RLS et SARIMA + AEKF pourrait être envisagée afin de tirer parti des forces complémentaires de ces approches. De plus, l'exploration d'autres techniques basées sur l'apprentissage machine pourrait également enrichir les capacités prédictives pour des contextes de trafic encore plus complexes.

En définitive, ce travail offre une contribution significative à la prédiction de flux routier, en mettant en évidence l'importance de choisir le bon modèle en fonction des contraintes du système à modéliser. Les résultats de cette étude pourraient également servir de base à l'optimisation des systèmes de gestion du trafic, en favorisant une prise de décision plus éclairée et efficace.

Dans le volet de la sécurité, l'étude a introduit une approche basée sur l'apprentissage automatique pour l'évaluation en temps réel des risques aux intersections, utilisant la

détection d'anomalies DBSCAN et un modèle de risque GPD pour classifier les niveaux de sécurité des intersections. Contrairement aux méthodes traditionnelles basées sur les données historiques d'accidents, notre approche évalue dynamiquement des mesures de sécurité substitutives, telles que les anomalies de PET et de vitesse, afin d'améliorer la gestion du trafic et la prévention des accidents.

Une limitation de notre approche réside dans l'utilisation de paramètres fixes de DBSCAN pour toutes les intersections, ce qui peut réduire son efficacité à détecter les anomalies dans des conditions de trafic variées. Pour remédier à cela, nous suggérons un modèle de détection d'anomalies unifié qui s'ajuste dynamiquement aux différents profils d'intersection en utilisant un seuil bivarié généralisé. De plus, pour améliorer la détection d'événements rares, DBSCAN pourrait être combiné avec une méthode d'apprentissage automatique adaptative qui optimise les paramètres de regroupement en temps réel.

Le GPD modélise efficacement la distribution globale des excédents de PET et de vitesse, offrant des informations précieuses pour l'évaluation des risques. Cependant, ses performances peuvent être moins fiables lorsqu'il est appliqué à une seule intersection.

En termes de performance de classification, le réseau de neurones optimisé (ONN) a surpassé les autres modèles, notamment dans l'identification des scénarios à haut risque (rouge) et à bas risque (vert). Le modèle de Bayes naïf gaussien (GNB) a montré une précision raisonnable pour la classification des risques modérés (jaune), mais ONN est resté le plus efficace globalement, atteignant la meilleure précision globale. Des améliorations supplémentaires, telles que des techniques d'équilibrage des données et l'ingénierie des caractéristiques, pourraient améliorer la robustesse de la classification et sa généralisation aux conditions réelles.

Ces résultats soulignent le potentiel de l'apprentissage automatique dans l'analyse de la sécurité routière et fournissent un cadre précieux pour intégrer l'évaluation en temps réel des risques dans les systèmes de gestion du trafic urbain. L'ONN est le meilleur parmi les quatre modèles pour classifier les catégories *green* et *red*, tandis que le GNB est relativement bon pour classifier la catégorie *yellow* selon la validation et les tests, mais l'ONN reste le meilleur en termes de précision globale.

Ce mémoire s'est concentré sur l'amélioration de la gestion des ressources municipales et l'optimisation des infrastructures routières en utilisant des approches modernes et innovantes. Les objectifs initiaux étaient d'améliorer l'efficacité du déploiement des ressources pour rendre les routes plus sécuritaires, fluides et écologiques, d'appliquer des techniques émergentes comme l'intelligence artificielle dans le domaine du transport, d'identifier des patrons et marqueurs pertinents pour la sécurité et la fluidité, et de prédire les niveaux de risque ou les volumes d'usagers.

À travers l'étude et l'implémentation des méthodes présentées, nous avons démontré que l'utilisation de techniques comme les modèles prédictifs SARIMA, SARIMA + AEKF et RLS permet non seulement d'améliorer la précision des prédictions de flux routier, mais également de fournir des outils pratiques pour répondre aux défis dynamiques du trafic urbain. Ces approches contribuent à optimiser l'utilisation des ressources, à réduire les risques liés à la sécurité, et à anticiper les volumes d'usagers dans un horizon temporel rapproché.

Les résultats obtenus montrent une avancée significative dans l'application de l'IA au domaine de l'ingénierie routière, notamment par la capacité à identifier des patrons clés dans l'activité quotidienne sur les artères de la ville. Cependant, il subsiste des défis, comme

la prise en compte d'événements imprévus ou de données manquantes, qui ouvrent la voie à des recherches futures.

Enfin, ce travail a démontré que l'intégration de technologies émergentes peut transformer la gestion des infrastructures urbaines en améliorant la sécurité, la fluidité et l'efficacité écologique des systèmes routiers. Ces résultats peuvent servir de base pour développer de nouvelles approches dans le domaine du transport intelligent et durable.

Améliorations et directions futures

Les travaux réalisés dans ce mémoire ouvrent la voie à plusieurs améliorations et directions futures qui pourraient enrichir les résultats obtenus et leur applicabilité dans le domaine du transport et de l'ingénierie routière.

Améliorations potentielles

Intégration de modèles prédictifs avancés : Explorer des approches telles que les réseaux neuronaux récurrents (RNN), les modèles LSTM (Long Short-Term Memory) ou les forêts aléatoires pour améliorer la précision et la robustesse des prédictions.

Prise en compte des anomalies : Développer des mécanismes pour détecter et gérer efficacement les anomalies dans les données, telles que les accidents ou les événements imprévus, afin de garantir la fiabilité des prédictions [25].

Amélioration de la collecte de données : Utiliser des capteurs modernes et des technologies IoT (Internet des objets) pour une collecte en temps réel de données plus complètes, incluant des facteurs comme les conditions météorologiques ou les comportements des usagers [26].

Optimisation des ressources municipales : Effectuer des simulations pour tester l'efficacité des prédictions dans la gestion proactive des ressources, comme les interventions liées à la fluidité ou à la sécurité routière.

Directions futures

Développement de systèmes de transport intelligents : Intégrer les modèles prédictifs dans des systèmes de gestion automatisée du trafic, permettant des ajustements en temps réel pour améliorer la fluidité et la sécurité.

Approche écologique : Étudier l'impact environnemental des embouteillages et proposer des solutions basées sur les prédictions pour réduire les émissions de gaz à effet de serre.

Analyse comparative avec d'autres régions : Effectuer une comparaison des résultats obtenus avec des données provenant d'autres villes ou régions, afin d'évaluer l'applicabilité des modèles dans divers contextes.

Collaboration multidisciplinaire : Impliquer des experts en urbanisme, écologie et sécurité routière pour enrichir les perspectives et favoriser des solutions intégrées et durables.

Ces améliorations et directions futures permettront d'approfondir l'impact de ce travail, tout en contribuant à l'évolution des systèmes de gestion du trafic vers des solutions plus intelligentes, durables et adaptées aux besoins contemporains.

Références

[1] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis Prevention*, vol. 38, no. 4, pp. 811–822, 2006.

[2] A. Boyle and C. O'Flaherty, *Highways, Fourth Edition*. Taylor & Francis, 2002.

[3] M. Eskandari Torbaghan, M. Sasidharan, L. Reardon, and L. C. Muchanga-Hvelplund, "Understanding the potential of emerging digital technologies for improving road safety," *Accident Analysis and Prevention*, vol. 166, 2022.

[4] E. Allie, "Mesures de la congestion routière à partir de données sur la vitesse instantanées de l'Étude sur l'utilisation des véhicules au canada," *Transport Canada*, 2016.

[5] P. Dong and Q. Chen, *LiDAR Remote Sensing and Applications*, ser. Remote Sensing Applications Series. CRC Press, 2017.

[6] T.Fall, D.Massicotte, J.Dessureault, and M.Ouameur, "Urban intersection safety risk index: Machine learning methods for real-time classification," *DSP2025 Costa Navarino*, 2025.

[7] P. Mcmanamon, *LiDAR Technologies and Systems*. SPIE, Jul. 2019, chapitre 4, p.132. ISBN: 9781510625396.

[8] R. A. Vincent and M. Ecker, "Light detection and ranging (lidar) technology evaluation." Oct 2010, tech Report.

[9] A. Stathopoulos, L. Dimitriou, and T. Tsekeris, "Fuzzy modeling approach for combined forecasting of urban traffic flow," *Computer-Aided Civil and Infrastructure Engineering*, vol. 23, no. 7, pp. 521–535, 2008.

[10] L. Cai, Z. Zhang, J. Yang, Y. Yu, T. Zhou, and J. Qin, "A noise-immune kalman filter for short-term traffic flow forecasting," *Physica A: Statistical Mechanics and its Applications*, vol. 536, p. 122601, 2019.

[11] J. Guo, W. Huang, and B. M. Williams, "Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014, special Issue on Short-term Traffic Flow Forecasting.

[12] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[13] National Research Council, *Highway Safety Manual*. American Association of State Highway and Transportation Officials, 2010, no. vol. 1, iSBN: 978-1-56051-477-0, LCCN: 2012471379.

[14] Y. Hu, Y. Li, C. Yuan, and H. Huang, "Modeling conflict risk with real-time traffic data for road safety assessment: a copula-based joint approach," *Transportation Safety and Environment*, vol. 4, no. 3, p. tdac017, Aug. 2022.

[15] A. Arun, M. M. Haque, S. Washington, T. Sayed, and F. Mannering, "A systematic review of traffic conflict-based safety measures with a focus on application context," *Analytic Methods in Accident Research*, vol. 32, p. 100185, 2021.

[16] S. Zhang and M. Abdel-Aty, "Real-time pedestrian conflict prediction model at the signal cycle level using machine learning models," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 176–186, 2022.

[17] R. Ezzati Amini, K. Yang, and C. Antoniou, "Development of a conflict risk evaluation model to assess pedestrian safety in interaction with vehicles," *Accident Analysis Prevention*, vol. 175, p. 106773, 2022.

[18] B. Yu, Y. Wang, Q. Chen, X. Chen, Y. Zhang, K. Luan, and X. Ren, "A review of road 3d modeling based on light detection and ranging point clouds," *Journal of Road Engineering*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274453294

[19] B. Everitt, S. Landau, M. Leese, D. Stahl, and a. O. M. C. Safari, *Cluster Analysis, 5th Edition*. John Wiley & Sons, 2011.

[20] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, Jul 1964.

[21] A. Carianni and A. Gemma, "Overview of traffic flow forecasting techniques," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 6, pp. 848–882, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:279485095

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[23] S. Coles, "An introduction to stat. modeling of extreme values," *Journal of the American Statistical Association*, vol. 97, pp. 1204–1204, 2002.

[24] MathWorks, "Design time series narx feedback neural networks," Natick, Massachusetts, United States, 2024.

[25] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[26] S. V. Ducca and C. B. Margi, "Performance trade offs in iot-based traffic monitoring and incident detection systems," in *2022 Symposium on Internet of Things (SIoT)*, 2022, pp. 1–4.

Annexe A - "Urban intersection risk index: Machine learning methods for real-time classification"

Urban intersection safety risk index: Machine learning methods for real-time classification

Thierno Fall
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
Mame.Thierno.Mbacke.Fall@uqtr.ca

Daniel Massicotte
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
Daniel.Massicotte@uqtr.ca

Jean-Sébastien Dessureault
Departement of Mathematics and Computer science
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
jean-sebastien.dessureault@uqtr.ca

Messaoud Ahmed Ouameur
Departement of Electrical and Computer Engineering
Université du Québec à Trois-Rivières
Trois-Rivières, Canada
messaoud.ahmed.ouameur@uqtr.ca

2025 25th International Conference on Digital Signal Processing (DSP) | 979-8-3315-1213-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/DSP65409.2025.11075008

***Abstract*—The safety of urban intersections is a critical concern for city planners. Technological advancements, such as LiDAR sensors, enable better risk assessment for road users. This study proposes a hybrid model that combines Post-Encroachment Time (PET) data with unsupervised machine learning techniques, specifically DBSCAN clustering, to detect traffic anomalies. A generalized Pareto distribution (GPD) is then applied to estimate a risk index. Finally, categorical safety risk classification is performed using an optimizable neural network (ONN), support vector machine (OSVM), efficient logistic regression (ELR), and Gaussian Naïve Bayes (GNB). The impact of these methods is evaluated in real-time for urban traffic management in Trois-Rivières, Quebec, Canada. This work aims to assist decision-makers in urban traffic planning and accident prevention.**

***Index Terms*—Surrogate measures, Post-Encroachment Time (PET), Machine learning, DBSCAN, Pareto distribution, Classification.**

I. Introduction

Intelligent transport systems develop and integrate methods to overcome the high demand for economic concerns, the reliability and quality of infrastructures, and most importantly, the safety of road users. Each year, 1.35 million people are killed on the roads of the world, and another 20 to 50 million are seriously injured [1].

In road traffic, many users interact with each other, and the need for reliable measurement systems (on a week, day, and even hour basis) of interactions between users is growing. These interactions will likely generate conflicts (between vehicles, pedestrians, and motorcycles) as they cross paths in all directions. Thus, conflicts occur when traffic streams moving in different directions interfere. The number of possible conflict points at any intersection depends on the number of approaches, the turning movements, and the type of traffic control at the intersection [2].

PET is calculated as when the first road user leaves the conflict point and the second reaches the same point. It is defined as the time between moment t_1 when the first road user exits the conflict point and moment t_2 when the second enters the same conflict spot, as shown in Figure 1. The smaller the PET, the higher the risk of collision. Moreover, a PET value less than zero would indicate a crash occurrence [3].

Due to the rarity of dangerous events, such as accidents and near-misses, and the lack of timeliness, having a reliable model that gives a safety index of road traffic is challenging. Surrogate traffic measures can be used to determine how dangerous an intersection or road section is.

[4] proposes a conflict-based traffic safety assessment method by associating conflict frequency and severity with short-term traffic characteristics. [5] conducts a systematic review of conflict-based safety measures with a specific focus on the context of their applications. [6] uses conflict indicators, PET and Time to Collision (TTC) to identify pedestrian conflicts and predict pedestrian conflicts one cycle ahead, which can be 2-3 min, whereas [7] employed surrogate safety indicators to measure the safety level of pedestrian conflict with other road users to evaluate conflict risk.

Despite the significant increase in interest in surrogate measures, many approaches lack real-time applicability. This study introduces a machine learning-based risk index capable of real-time classification of intersection safety levels. Our contribution uses PET as a surrogate measure and focuses on 1) machine learning techniques to evaluate the behavior of individual road intersections, 2) calculating the safety index for an hourly divided block of conflicts of all the data frames from ten intersections, and 3) using the generated safety risk level and conflict features (PET and speed) to train and test classification methods.

The paper is structured as follows: Section II presents the methodology, including data description, anomaly detection, the generalized Pareto Risk Index, and safety level classification. Section III discusses the results and their implications. Finally, conclusions are provided in Section IV.

II. Methodology

A. Data description

Traffic data are collected via Velodyne LiDAR sensors by third-party BlueCity. Thus, we assume the collected data completely satisfy the requirements of reliability, repeatability, and practicability [8]. The main fields from accessible conflict data are PET and speed (involved speed). The PET data range is from 0 to 10 seconds, whereas the speed can be up to 100 km/h. For the present contribution, PET and speed, more

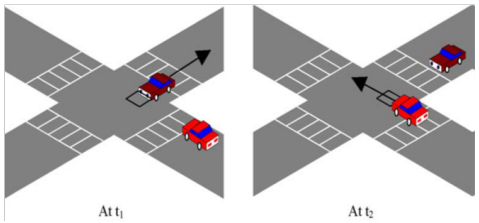


Fig. 1: Illustration of PET conflicts [3]

than 4 seconds and less than 20 km/h, respectively, are not considered. Concerning [3], we use the negated PET instead of the actual PET.

The conflicts are event-based and therefore remain stochastic. As depicted by Table I, the dataset comprises, in addition to PET and speed, date and time, types of users involved, and geographical locations to identify intersection hot spots.

B. Anomaly detection

The DBSCAN addresses the incompatibility of the classical cluster algorithm (i.e., K-means) to particular shaped data as stated in [9]. DBSCAN clustering is used to classify conflict points into normal and abnormal events. A density-based approach ensures the identification of rare, high-risk occurrences. After analyzing the architecture of our data, the intuitive notion of "clusters" and "noises" in a given database [9] is easily applicable. There is a large density of points around a range of PET and speed, which appear to be clustered and might be interpreted as normal or usual behavior in users' conflicts surrounded by rare and infrequent conflicts.

Serious conflicts leading to fatalities and general interactions between road users are less frequent than usual. It can be observed from the joint distribution shown in Figure 2 of PET and speed. Thus, these dangerous conflict data points appear to be noise with the DBSCAN technique according to the definition of the algorithm and the expected rarity of these events.

In this work, the threshold is not determined using the conventional parameter stability analysis method. Instead, it is defined as a percentile, computed from the ratio between the number of noisy points and the average number of points within the identified clusters. This ratio-based approach provides a data-driven and adaptive way to establish the threshold.

The parameters of the DBSCAN, clusters minimum points γ and maximal distance between points ϵ were determined empirically through a grid search to balance anomaly detection sensitivity and noise reduction. These values optimize the distinction between normal and anomalous traffic conflicts while maintaining high detection precision. That maneuver limits performance because intersection traffic activities are relatively different.

C. Generalized Pareto risk index

Exceedances of negated PET and speed are modeled using a GPD. Following the approach of [3], this study adheres closely to the formulations and theoretical foundations presented in [10]. The risk index is defined by Equation (1).

$$R_i x = Pr(X_i > \mu) = 1 - G(X_i) \tag{1}$$

Where $G(X_i)$ is the GPD function, μ is the threshold limit, and X_i is the random variable.

TABLE I: Conflicts fields informations

Feature	Description	Type	Range
PET	Time difference between two users crossing same point	Continuous	0 - 10 seconds
Speed	Speed difference between involved users	Continuous	$>$ 10 km/h
$1^{st} user$		Categorical	Car, Pedestrian, Bus, Bicycle, Motorcycle, Truck
$2^{nd} user$		Categorical	Car, Pedestrian, Bus, Bicycle, Motorcycle, Truck
Date and Time	Date and time of occurence	Date	
Position	GPS coordinates where the conflicts occurs	Decimal	Latitude and Longitude

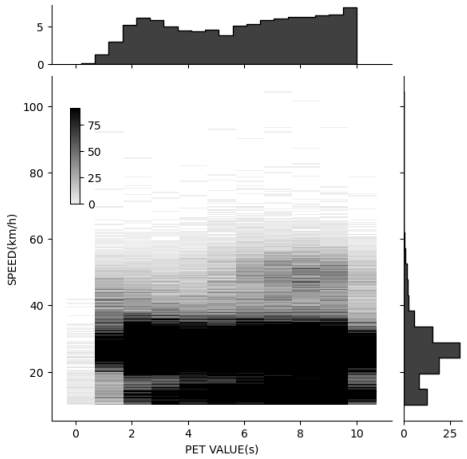


Fig. 2: Joint Marginal Histogram of PET and speed

As demonstrated by [10], the GPD function of (1) is formulated in (2).

$$G(x) = \exp\left(-\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right) \tag{2}$$

where the parameters are the shape, the scale and the threshold namely $\xi \neq 0$, σ and μ, respectively.

The choice of the GPD is justified by the exponential decay observed in the tail of our data distribution [11], particularly for the negated PET. The parameters of the GPD distribution are estimated using maximum likelihood estimation (MLE), with 95% confidence intervals for parameter estimates to ensure an optimal fit to the data. Subsequently, Equation (1) is applied to the PET and speed exceedance data for the i^{th} hourly block, computing $R_i p$ and $R_i s$ separately. The overall risk index R_i is then obtained as the mean of these two values. Once completed, a theoretical safety category is established based on the computed risk index.

$$R_i = \frac{R_i p + R_i s}{2} \tag{3}$$

D. Classification of the safety risk level

All previously collected PET and speed data are available, making it straightforward to compute the risk index as de-

scribed in II-C. For classification, the input is conflict features PET and speed. We employ four machine learning models: neural networks, support vector machines (SVM), efficient logistic regression (ELR), and Gaussian naïve Bayes (GNB). Among these, the neural network and SVM undergo further optimization. The optimized parameters for the neural network include: the number of fully connected layers between 1 - 3, the activation function: ReLU, tanh, sigmoid or none, the regularization strength between 3.31 10^{-10} - 3.31, the first and second layer sizes between 1-300 and standardized data or not. For the SVM model, the optimized parameters include: the kernel function: gaussian, linear, quadratic or cubic, the constraint level: 0.001 - 1000, the multiclass coding: One-vs-All or One-vs-One and standardized data or not.

The safety level classification is defined as follows: $green$ safety category if $R_i \leq 0.35$, $yellow$ if $0.35 < R_i \leq 0.65$ and if $R_i < 0.65$ then the category is red. Each conflict data point serves as an input for the classification model. The classification thresholds were empirically determined through sensitivity analysis to optimize model accuracy while ensuring meaningful risk differentiation.

III. Results and Discussion

The primary goal of this study is to accurately predict the safety level (category) of intersections in the city of Trois-Rivières. All simulations and implementations were carried out using MATLAB tools. For anomaly detection, the DBSCAN parameters are set to ϵ = 0.45 and γ = 220. As shown in Figure 3, the identified abnormal data points align with those in the joint distribution presented in Figure 2.

In Figure 4, the blue histogram represents the probability density of PET exceedances, capturing values below the threshold (in the context of negative values). The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line depicts the GPD fit applied to the exceedance data. The sharp increase near zero suggests that the distribution is heavily concentrated around low exceedance values, a characteristic commonly observed in distributions with Pareto-like tails.

In Figure 4a, the blue histogram represents the probability density of PET exceedances, capturing values below the threshold (in the context of negative values). The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line depicts the GPD fit applied to the exceedance data. The strong alignment between the histogram and the red curve in Figure 4a demonstrates a good fit, validating the use of the GPD to model PET exceedance data. This confirms the suitability of GPD for capturing extreme PET values.

Table II provides a summary of the fit results for both PET and speed. The shape parameter (ξ = -0.16) indicates a bounded tail, implying a limited range of extreme values. However, slight discrepancies between the fitted GPD and the histogram in Figure 4a arise due to data sparsity in extreme-value regions.

In Figure 4b, the blue histogram represents the probability density of speed exceedances. The bars indicate the frequency of these exceedances, normalized to reflect a probability density. The red line shows the GPD fit applied to the speed exceedance data. The positive shape parameter (ξ = 0.19) indicates a heavy-tailed distribution, suggesting a higher probability of extreme values.

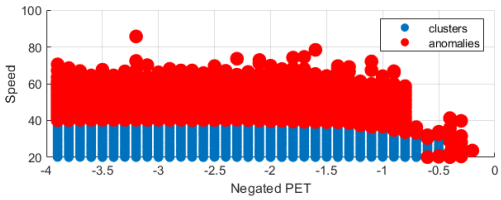
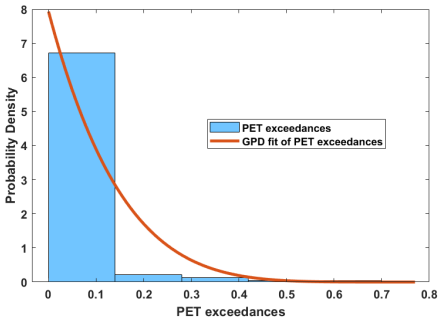
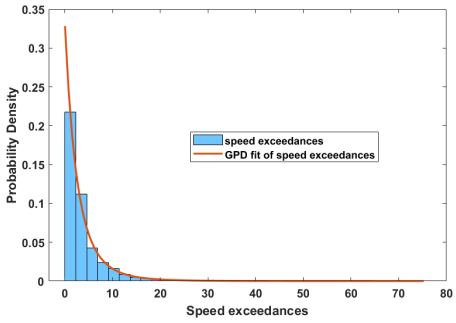


Fig. 3: DBSCAN anomaly detection



(a) Fit of Negated PET



(b) Fit of speed

Fig. 4: GPD Fit of PET and Speed Exceedances

For classification, 75% of the data is allocated for training and validation, with cross-validation applied to mitigate overfitting. The remaining 25% is reserved for testing. The training dataset consists of a total of 41079 samples, distributed across classes. The class distribution is as follows: Class $green$ – 5720 samples, Class $yellow$ – 28084 samples, and Class red - 7275 samples. As shown in Table III, overall accuracy is used as a key performance metric, where higher accuracy indicates better classification performance. Additionally, total cost represents the number of misclassified instances (lower is better), while the error rate serves as an indicator of the model's misclassifications tendency.

Further insights can be gained from metrics such as the true positive rate (TPR), which quantifies the proportion of correctly predicted instances within each class, and the false negative rate (FNR), which measures the proportion of misclassified negative instances. The confusion matrices presented in Tables IV detail the classification performance for test datasets across different methods.

The comparative classification is conducted using four models: optimizable neural network (ONN), optimizable support vector machine (OSVM), efficient logistic regression (ELR), and Gaussian naïve Bayes (GNB). The evaluation metrics include accuracy (for test datasets), TPR, FNR, and confusion matrices to assess model performance.

TABLE II: GPD Fit Parameters

Negated PET GPD fit			Speed GPD fit		
Shape:ξ	*Scale:*σ	*Thres.:*μ	*Shape:*ξ	*Scale:*σ	*Thres.:*μ
-0.16	0.12	-0.9	0.19	3.04	56.2

TABLE III: Test and validation results

	Validation			Test		
	Acc.	**T.Cost**	**E.Rate**	**Acc.**	**T.Cost**	**E.Rate**
ONN	*78.95%*	*6340*	*21.05%*	*79.37%*	*5084*	*20.63%*
OSVM	*77.69%*	*6720*	*22.31%*	*78.88%*	*5205*	*21.12%*
ELR	*76.82%*	*6983*	*23.18%*	*78.14%*	*5387*	*21.86%*
GNB	*75.82%*	*7285*	*24.18%*	*77.72%*	*5491*	*22.28%*

The majority of PET and speed data points fall into the $yellow$ category, indicating a high classification accuracy for this class, as shown in the confusion matrix results in Tables IV. To address this imbalance, the dataset should be adjusted to ensure a more even distribution across all classes. Among the tested models, optimized neural network (ONN) achieved the highest accuracy, reaching 78.95% for validation and 79.37% for testing. It was followed by the optimized support vector machine (OSVM) with 77.69% and 78.88%, respectively, then ELR 76.82% and 78.14% and finally GNB 75.82% and 77.72%. The optimal parameters for the ONN model include: 2 fully connected layers, a tanh activation function, a regularization strength of $4.10e^{-10}$, a first layer size of 298, a second layer size of 3 and standardized data. For OSVM, the optimized parameters are: a quadratic kernel function, a constraint level of 39.76, One-vs-One multiclass coding, and standardized input data. ONN outperforms other models due to its ability to capture non-linear relationships, extract hierarchical features, and optimize hyperparameters for better adaptability. Its robustness in classifying both high-risk (red) and low-risk (green) scenarios makes it the most reliable choice for real-time risk assessment.

Some may argue that the GPD alone is sufficient for classifying conflicts. However, its limitations must be considered, particularly its reliance on complete data blocks over a set period before classification can be performed. While the classification methods in this study use instantaneous PET and speed as predictors (bi-variate input), they operate on individual data points rather than aggregated groups, enabling real-time processing. Moreover, their accuracy is still evaluated based on the GPD framework (theoretical classification of computed R_i).

IV. Conclusion

This study introduced a machine learning-based approach for real-time intersection risk assessment, leveraging DBSCAN anomaly detection and a GPD risk model to classify intersection safety levels. Unlike traditional methods based on historical accident data, our approach dynamically evaluates surrogate safety measures, such as PET and speed anomalies, to improve traffic management and accident prevention.

One limitation of our approach is the use of fixed DBSCAN parameters across all intersections, which may reduce its effectiveness in detecting anomalies in varying traffic conditions. To address this, we suggest a unified anomaly detection model that dynamically adjusts to different intersection profiles using a generalized bi-variate threshold. Additionally, to improve the detection of rare events, DBSCAN could be combined with an adaptive machine learning method that optimizes clustering parameters in real-time.

TABLE IV: Test confusion matrix results

ONN		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2945(**71.8%**)*	*1072(26.1%)*	*83(2%)*	***28.2%***
	Yellow	*980(7%)*	*12229(86.8%)*	*886(6.3%)*	*13.2%*
	Red	*204(3.2%)*	*1859(28.8%)*	*4388(**68%**)*	***32%***
OSVM		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2444(59.6%)*	*1605(39.1%)*	*51(1.2%)*	*40.4%*
	Yellow	*501(3.6%)*	*13051(92.6%)*	*543(3.9%)*	*7.4%*
	Red	*145(2.2%)*	*2360(36.6%)*	*3946(61.2%)*	*38.8%*
ELR		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2643(64.5%)*	*1438(35.1%)*	*19(0.5%)*	*35.5%*
	Yellow	*648(4.6%)*	*12994(92.2%)*	*453(3.2%)*	*7.8%*
	Red	*199(3.1%)*	*2630(40.8%)*	*3622(56.1%)*	*43.9%*
GNB		**Predicted**			
		Green	**Yellow**	**Red**	**FNR**
True	**Green**	*2321(56.6%)*	*1740(42.4%)*	*39(1%)*	*43.4%*
	Yellow	*435(3.1%)*	*13226(**93.8%**)*	*434(3.1%)*	***6.2%***
	Red	*174(2.7%)*	*2669(41.4%)*	*3608(55.9%)*	*44.1%*

The GPD effectively models the overall distribution of PET and speed exceedances, offering valuable insights for risk assessment. However, its performance may be less reliable when applied to a single intersection.

In terms of classification performance, the optimized neural network (ONN) outperformed other models, particularly in identifying high-risk (red) and low-risk (green) scenarios. Gaussian Naïve Bayes (GNB) demonstrated fair accuracy for moderate-risk (yellow) classification, but ONN remained the most effective overall, achieving the highest global accuracy. Further refinements, such as data balancing techniques and feature engineering, could enhance classification robustness and generalization to real-world conditions.

These findings highlight the potential of machine learning in traffic safety analysis and provide a valuable framework for integrating real-time risk assessment into urban traffic management systems. ONN is the best among the four to classify $green$ and red, GNB is fair to classify $yellow$ according to test, but ONN is the best according to global accuracy.

Acknowledgment

This work has been funded by the Natural Sciences and Engineering Research Council of Canada, Canada Foundation for Innovation, Ville de Trois-Rivières, and the Research Chair in Signals and Intelligence of High-Performance Systems.

REFERENCES

[1] M. Eskandari Torbaghan, M. Sasidharan, L. Reardon, and L. C. Muchanga-Hvelplund, "Understanding the potential of emerging digital technologies for improving road safety," *Accident Analysis Prevention*, vol. 166, p. 106543, 2022.

[2] A. Boyle and C. O'Flaherty, *Highways, Fourth Edition*. Taylor & Francis, 2002.

[3] P. Songchitruksa and A. P. Tarko, "The extreme value theory approach to safety estimation," *Accident Analysis Prevention*, vol. 38, no. 4, pp. 811–822, 2006.

[4] Y. Hu, Y. Li, C. Yuan, and H. Huang, "Modeling conflict risk with real-time traffic data for road safety assessment: a copula-based joint approach," *Transportation Safety and Environment*, vol. 4, no. 3, p. tdac017, 08 2022. [Online]. Available: https://doi.org/10.1093/tse/tdac017

[5] A. Arun, M. M. Haque, S. Washington, T. Sayed, and F. Mannering, "A systematic review of traffic conflict-based safety measures with a focus on application context," *Analytic Methods in Accident Research*, vol. 32, p. 100185, 2021.

[6] S. Zhang and M. Abdel-Aty, "Real-time pedestrian conflict prediction model at the signal cycle level using machine learning models," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 176–186, 2022.

[7] R. Ezzati Amini, K. Yang, and C. Antoniou, "Development of a conflict risk evaluation model to assess pedestrian safety in interaction with vehicles," *Accident Analysis Prevention*, vol. 175, p. 106773, 2022.

[8] W. D. Glauz and D. J. Migletz, "Application of traffic conflict analysis at intersections," *NCHRP Report*, 1980.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[10] S. Coles, "An introduction to stat. modeling of extreme values," *Journal of the American Statistical Association*, vol. 97, 2001.

[11] T. M. Inc., "Design time series narx feedback neural networks," Natick, Massachusetts, United States, 2024.