



Titre: Développement d'un outil prédictif de la génération de résidus dans un contexte de logistique inverse

Auteur: Aicha Ksouma

Date: 2026

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Ksouma, A. (2026). Développement d'un outil prédictif de la génération de résidus dans un contexte de logistique inverse [Master's thesis, Polytechnique Montréal]. PolyPublie. <https://publications.polymtl.ca/76916/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/76916/>

Directeurs de recherche: Samira Keivanpour, & Jean-Marc Frayret

Programme: Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Développement d'un outil prédictif de la génération de résidus dans un
contexte de logistique inverse**

AICHA KSOUMA

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie industriel

Mai 2026

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Développement d'un outil prédictif de la génération de résidus dans un
contexte de logistique inverse**

présenté par **Aicha KSOUMA**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Martin TRÉPANIER, président

Samira KEIVANPOUR, membre et directrice de recherche

Jean-Marc FRAYRET, membre et codirecteur de recherche

Maha BEN ALI, membre

DÉDICACE

*“Chercher, c’est accepter de ne pas savoir.
Et peut-être que la vie commence précisément là,
dans cet espace fragile où l’on avance sans certitude.”*

— Eni AK

Je dédie ce mémoire à ma chère famille,
pour leur amour inconditionnel, leur soutien indéfectible et leurs sacrifices silencieux qui ont
rendu ce parcours possible.

À mon père, Nawfel, pour sa force tranquille, sa confiance constante et les valeurs qu’il m’a
transmises. Pour m’avoir appris la persévérance, le courage face aux épreuves et
l’importance de croire en soi, même lorsque le chemin semble incertain.

À ma mère, Hayet, qui m’a appris que la persévérance naît du cœur et que rien n’est
impossible lorsque l’on avance avec foi.

À ma tante Ammatou, ma seconde mère et ma professeure de français à temps plein, pour
sa présence rassurante, sa bienveillance infinie et ses corrections discrètes qui m’ont appris à
aimer la précision des mots autant que la vie.

À mes amis qui ont cru en moi, y compris dans les moments de doute, et qui m’ont
encouragée à continuer, à avancer et à ne jamais abandonner.

Ce travail est autant le fruit de votre confiance que de mes propres efforts.

REMERCIEMENTS

La rédaction de ce mémoire constitue l’aboutissement de mon parcours de maîtrise en génie industriel à Polytechnique Montréal. Ce travail n’aurait pas pu voir le jour sans le soutien, les conseils et l’accompagnement de nombreuses personnes que je souhaite remercier sincèrement.

Je tiens d’abord à exprimer ma profonde gratitude à ma directrice de recherche, la professeure Samira Keivanpour, pour la confiance qu’elle m’a accordée, sa grande disponibilité et la qualité de ses conseils tout au long de ce projet. Son encadrement rigoureux, sa vision scientifique et son soutien constant m’ont permis de progresser tant sur le plan académique que personnel.

J’adresse également mes remerciements à mon co-directeur, le professeur Jean-Marc Frayret, pour ses orientations stratégiques, ses remarques toujours pertinentes et la perspective industrielle qu’il a apportée, contribuant ainsi à enrichir significativement cette recherche.

Je souhaite aussi remercier Ashkan Arminia pour son aide technique, la richesse de nos échanges scientifiques et son soutien à chacune des étapes du projet.

Ma reconnaissance va également à nos entreprises partenaires Oscar AI et RPM Eco, pour l’accès aux données industrielles et pour la collaboration qui a permis d’ancrer ce travail dans des enjeux concrets et réels de terrain.

Je souhaite également remercier le programme Mitacs Accelerate pour son soutien financier et pour avoir rendu possible cette collaboration entre le milieu universitaire et le milieu industriel.

Je remercie en outre mes collègues de laboratoire et amis pour leur entraide, nos discussions stimulantes et l’ambiance bienveillante qui a marqué ces années d’études.

Enfin, j’exprime mes remerciements les plus chaleureux à ma famille pour son amour, sa patience et ses encouragements constants. Leur soutien indéfectible a été ma principale source de motivation tout au long de ce parcours.

À toutes celles et tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire, j’adresse l’expression de ma sincère gratitude.

RÉSUMÉ

La gestion des résidus industriels représente un enjeu majeur pour les systèmes de logistique inverse et les stratégies d'économie circulaire. La planification efficace des opérations de collecte repose notamment sur la capacité à anticiper les volumes générés. Toutefois, la prévision de ces flux demeure complexe en raison de leur forte variabilité temporelle, de leur hétérogénéité spatiale et du caractère intermittent des séries observées.

Ce mémoire propose un cadre méthodologique fondé sur l'apprentissage automatique afin d'améliorer la prévision des quantités de résidus industriels. L'étude s'appuie sur des données réelles de collecte de bidons d'huile de moteur usagés au Canada, couvrant la période 2014–2024 et comprenant plus d'un million d'observations.

Le travail est structuré en deux articles complémentaires. Le premier article analyse l'impact des stratégies d'agrégation spatio-temporelle sur la performance des modèles prédictifs. Deux niveaux d'agrégation spatiale sont comparés : une agrégation au niveau des villes et une partition fine basée sur le système d'indexation hexagonale hiérarchique H3. Plusieurs modèles sont évalués, incluant des approches statistiques (ARIMA, Croston) et des méthodes d'apprentissage automatique basées sur le gradient boosting (CatBoost, XGBoost, LightGBM).

Le second article propose une approche de lissage spatio-temporel fondée sur la similarité comportementale entre unités spatiales afin d'améliorer la robustesse des prévisions à haute résolution. Des variables météorologiques ont également été explorées comme facteurs explicatifs potentiels. Toutefois, leur contribution à l'amélioration des performances prédictives s'est révélée limitée dans le contexte étudié, suggérant que la génération des résidus est davantage influencée par des facteurs opérationnels et comportementaux. Le mécanisme proposé combine les prédictions initiales du modèle CatBoost avec une agrégation pondérée des prédictions des hexagones voisins présentant des profils similaires.

Les résultats expérimentaux montrent que les modèles d'apprentissage automatique surpassent significativement les approches statistiques traditionnelles. Ils mettent également en évidence un compromis entre précision prédictive et granularité spatiale. L'approche de lissage proposée permet de réduire les erreurs extrêmes et d'améliorer la stabilité des prédictions dans les zones à forte variabilité.

Les contributions de ce mémoire résident dans l'analyse systématique de l'effet de la granularité spatiale, l'intégration d'une indexation hiérarchique (H3), et la proposition d'un mécanisme de lissage spatio-temporel fondé sur la similarité. Ce travail apporte ainsi un cadre

méthodologique intégré permettant d'améliorer la prévision et la planification des systèmes de logistique inverse en contexte industriel réel.

ABSTRACT

Industrial waste management represents a critical challenge for reverse logistics systems and circular economy strategies. Efficient planning of collection operations relies on the ability to accurately forecast generated waste volumes. However, predicting these flows remains complex due to their high temporal variability, spatial heterogeneity, and intermittent nature.

This thesis proposes a machine learning-based methodological framework to improve the forecasting of industrial waste generation. The study is based on real-world data of empty motor oil container collections across Canada, covering the period 2014–2024 and comprising over one million observations.

The thesis is structured around two complementary research articles. The first article investigates the impact of spatio-temporal aggregation strategies on predictive performance. Two spatial aggregation levels are compared: city-level aggregation and a fine-grained partition based on the hierarchical hexagonal spatial indexing system H3. Several forecasting models are evaluated, including traditional statistical approaches (ARIMA and Croston) and gradient boosting methods (CatBoost, XGBoost, and LightGBM).

The second article introduces a spatial smoothing approach based on behavioral similarity between spatial units to enhance prediction robustness at high spatial resolution. Weather variables were also examined as potential explanatory factors. However, their contribution to improving predictive performance proved to be limited in the context of this study, suggesting that the generation of residuals is more strongly influenced by operational and behavioral factors. The proposed method combines initial CatBoost predictions with a weighted aggregation of predictions from neighboring hexagons exhibiting similar generation patterns.

Experimental results demonstrate that machine learning models significantly outperform traditional statistical methods. They also highlight a trade-off between predictive accuracy and spatial granularity. The proposed smoothing approach reduces extreme forecasting errors and improves prediction stability in highly variable regions.

The main contributions of this work include a systematic analysis of spatial granularity effects, the integration of hierarchical spatial indexing (H3), and the development of a similarity-based spatio-temporal smoothing mechanism. Overall, this research provides a comprehensive framework for improving forecasting and operational planning in industrial reverse logistics systems.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
RÉSUMÉ	v
ABSTRACT	vii
LISTE DES TABLEAUX	xii
LISTE DES FIGURES	xiii
LISTE DES SIGLES ET ABRÉVIATIONS	xiv
CHAPITRE 1 INTRODUCTION	1
1.1 Contexte général	1
1.2 Contexte spécifique	1
1.2.1 Énoncé du problème et lacunes	3
1.2.2 Contributions à la recherche	4
1.3 Plan de Mémoire	4
CHAPITRE 2 REVUE DE LITTÉRATURE	6
2.1 Méthodologie de recherche documentaire	6
2.2 Logistique inverse et génération des résidus industriels	6
2.2.1 Logistique inverse et économie circulaire	6
2.2.2 Variabilité et spécificité de la génération des résidus	7
2.2.3 Rôle de la prévision dans la performance des systèmes inverses	8
2.3 Prévision en logistique inverse et demande intermittente	9
2.3.1 Caractéristiques de la demande intermittente	9
2.3.2 Limites des approches classiques en contexte industriel	9
2.3.3 Apport des méthodes d'apprentissage automatique	10
2.4 Dimension spatiale en prévision et agrégation géographique	11
2.4.1 Intégration de la dimension spatiale en prévision	11
2.4.2 Problème d'agrégation spatiale et MAUP	11
2.4.3 Indexation spatiale hiérarchique et système H3	12
2.4.4 Implications pour la prévision des résidus industriels	13
2.5 Lissage spatio-temporel et approches par diffusion	13

2.5.1	Dépendance spatiale et lissage	13
2.5.2	Modèles de diffusion sur graphes	14
2.5.3	Similarité comportementale et voisinage adaptatif	14
2.5.4	Limites actuelles et opportunités de recherche	15
2.6	Clustering spatio-temporel et regroupement dynamique	15
2.6.1	Clustering de séries temporelles : fondements théoriques	15
2.6.2	Clustering spatial et segmentation géographique	15
2.6.3	Clustering dynamique et adaptation temporelle	16
2.7	Synthèse de la revue de la littérature	16
2.8	Lacunes et limites des approches existantes	18
2.9	Positionnement scientifique du mémoire	19
CHAPITRE 3 CADRE MÉTHODOLOGIQUE GÉNÉRAL		21
3.1	Objectifs de recherche	21
3.2	Méthodologie générale	21
3.3	Structure des articles	22
3.3.1	Article 1 : Préviation de la génération des résidus industriels	22
3.3.2	Article 2 : Amélioration des prévisions par lissage spatio-temporel	23
3.3.3	Préparation et enrichissement des données	24
3.4	Justification des choix méthodologiques	25
3.4.1	Choix des modèles d'apprentissage automatique	25
3.4.2	Justification du système d'indexation H3	26
3.4.3	Choix de l'agrégation temporelle	26
3.4.4	Mécanisme du lissage spatio-temporel	26
3.5	Présentation du cas d'étude et des données	26
CHAPITRE 4 ARTICLE 1 : OPTIMIZING REVERSE LOGISTICS : A SPATIO- TEMPORAL ML FRAMEWORK FOR INDUSTRIAL WASTE FORECASTING		30
4.1	Introduction	30
4.2	Literature Review	31
4.2.1	Reverse Logistics and Industrial Waste Generation	32
4.2.2	Waste Generation Forecasting in Reverse Logistics	32
4.2.3	Machine Learning and Advanced Forecasting Approaches	32
4.2.4	Geographical Clustering and Operational Gaps	33
4.2.5	Research Gap and Positioning of This Paper	34
4.3	STRUCTURE AND CONTEXT	34
4.3.1	Industrial Context and Partners	34

4.3.2	Datasets Description	35
4.3.3	Data Characteristics and Challenges	36
4.3.4	Study Area	36
4.4	Methodology	36
4.4.1	Overview of the Proposed Framework	36
4.4.2	Data cleaning and preprocessing	37
4.4.3	Temporal Aggregation Strategy	38
4.4.4	Spatial Aggregation Strategies	38
4.4.5	Feature Construction	40
4.4.6	Forecasting Models	42
4.4.7	Training and Evaluation Protocol	42
4.5	Results and Discussions	43
4.5.1	City-Level Forecasting Results	43
4.5.2	H3-Based Spatial Forecasting Results	43
4.5.3	Comparative Analysis and Operational Insights	45
4.6	Conclusion	45
CHAPITRE 5 ARTICLE 2 : DYNAMIC SPATIAL SMOOTHING BASED ON PRO-		
DUCTION PROFILE SIMILARITY FOR INDUSTRIAL WASTE FORECASTING		46
5.1	Introduction	46
5.2	Literature Review	48
5.3	Methodology	50
5.3.1	Proposed Similarity-Based Forecasting Framework	50
5.3.2	Weather Feature Integration	52
5.3.3	Dynamic Behavioral Profile Construction	52
5.3.4	Similarity Computation and Spatial Smoothing	53
5.3.5	Similarity-Based Spatial Smoothing	54
5.3.6	Case Study : Industrial Oil Container Recovery in Canada	55
5.4	Results and Discussion	56
5.4.1	Experimental Comparison of Forecasting Configurations	56
5.4.2	Baseline Forecasting Performance	57
5.4.3	Impact of Weather Variables	57
5.4.4	Impact of Similarity-Based Spatial Smoothing	58
5.5	Conclusion	60
CHAPITRE 6 CONCLUSION ET DISCUSSION GÉNÉRALE		61
6.1	Synthèse des travaux	61

6.2	Implications de la recherche	62
6.3	Limites	63
6.4	Recherches futures	64
RÉFÉRENCES		66

LISTE DES TABLEAUX

Table 2.1	Synthèse de la revue de littérature	17
Table 4.1	Summary of engineered features used for waste generation forecasting	41
Table 4.2	Performance comparison of forecasting models (City, biweekly)	43
Table 4.3	Performance comparison of forecasting models (H3, biweekly)	44
Table 5.1	Summary of spatial forecasting and similarity-based smoothing approaches	49
Table 5.2	Predictive performance comparison across model configurations . . .	56
Table 5.3	Percentile absolute error comparison	57

LISTE DES FIGURES

Figure 3.1	Cadre conceptuel du projet de recherche et articulation des deux articles.	22
Figure 3.2	Pipeline méthodologique de l'Article 1 : structuration spatio-temporelle et modélisation prédictive.	23
Figure 3.3	Approche proposée dans l'Article 2 : lissage spatio-temporel basé sur la similarité.	24
Figure 3.4	Extrait de la base de données <i>PickUp-PlasticContainer-Oil</i>	27
Figure 3.5	Extrait de la base de données <i>CollectionSites-PlasticContainer-Oil</i>	28
Figure 4.1	Spatial distribution of industrial collection sites across Eastern Canada.	35
Figure 4.2	Overall methodological workflow adopted in this study.	38
Figure 4.3	Comparison of waste generation patterns under different temporal aggregation levels (weekly, biweekly, and monthly) for the Toronto area	39
Figure 4.4	Hexagonal spatial aggregation of collection sites using the H3 indexing system.	40
Figure 5.1	Overview of the proposed dynamic similarity-based smoothing framework.	51
Figure 5.2	Illustration of the H3 k -ring neighborhood structure.	54
Figure 5.3	Spatial distribution of high-activity H3 hexagons for a representative bi-weekly period in Southern Ontario, Canada	56
Figure 5.4	Time-series example (Part 1) for a highly volatile hexagonal unit . . .	59
Figure 5.5	Time-series example (Part 2) for a highly volatile hexagonal unit . . .	59

LISTE DES SIGLES ET ABRÉVIATIONS

AI	Artificial Intelligence
ARIMA	AutoRegressive Integrated Moving Average
CatBoost	Categorical Boosting
CV	Cross-Validation
ECCC	Environment and Climate Change Canada
GNN	Graph Neural Network
H3	Hierarchical Hexagonal Spatial Indexing System
KPI	Key Performance Indicator
LightGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAUP	Modifiable Areal Unit Problem
ML	Machine Learning
NN	Neural Network
R^2	Coefficient de détermination
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
SVRP	Stochastic Vehicle Routing Problem
XGBoost	Extreme Gradient Boosting

CHAPITRE 1 INTRODUCTION

1.1 Contexte général

La transition vers des systèmes industriels plus durables constitue aujourd’hui un enjeu majeur pour les organisations et les pouvoirs publics. À l’échelle mondiale, la production de déchets solides a été estimée à environ 2,24 milliards de tonnes en 2020 et pourrait atteindre 3,88 milliards de tonnes d’ici 2050, illustrant l’ampleur croissante des pressions exercées sur les ressources, les infrastructures de gestion des déchets et l’environnement [1]. Au Canada, 36,5 millions de tonnes de déchets solides ont été générées en 2022, dont seulement 27,1 % ont été détournées de l’élimination, ce qui souligne l’importance d’améliorer les pratiques de récupération et de valorisation [2].

Dans ce contexte, l’économie circulaire occupe une place de plus en plus importante, puisqu’elle vise à prolonger le cycle de vie des produits, à favoriser la réutilisation et la valorisation des matières, et à réduire les volumes de déchets destinés à l’élimination [3], [4]. Contrairement au modèle linéaire traditionnel fondé sur la logique « extraire, produire, consommer, jeter », l’économie circulaire propose une gestion plus responsable des ressources en considérant les résidus comme des matières susceptibles d’être récupérées, recyclées ou réintégrées dans de nouveaux cycles de production.

Toutefois, la mise en œuvre concrète de ces principes repose sur la capacité à gérer efficacement les flux de retour, souvent complexes et incertains. C’est dans ce cadre que la logistique inverse joue un rôle central, en assurant la collecte, le transport et la valorisation des résidus industriels.

1.2 Contexte spécifique

La logistique inverse regroupe l’ensemble des activités permettant de ramener des produits, composants ou résidus depuis leur lieu d’utilisation vers des points de récupération, de recyclage, de remanufacturing ou d’élimination sécuritaire. À la différence de la logistique directe, les flux inverses sont marqués par une forte incertitude quant à la quantité, à la localisation, à la qualité et au moment d’apparition des retours [3]. Cette incertitude complique considérablement la planification des opérations, qu’il s’agisse du dimensionnement des ressources, de l’organisation des tournées ou du maintien d’un niveau de service satisfaisant à coût raisonnable.

Dans les environnements industriels, cette complexité est particulièrement marquée. Les volumes de résidus générés dépendent souvent de multiples facteurs, notamment les activités de maintenance, les rythmes d'exploitation, les pratiques locales, la saisonnalité et certains facteurs contextuels externes. Les flux de résidus industriels présentent ainsi une forte hétérogénéité spatio-temporelle : certaines zones génèrent des volumes relativement stables, tandis que d'autres présentent des comportements intermittents, irréguliers ou très variables.

Cette variabilité rend la gestion opérationnelle particulièrement difficile. Une mauvaise anticipation des volumes à collecter peut entraîner des déplacements inutiles, une sous-utilisation ou une saturation des capacités de transport, ainsi qu'une allocation inefficace des ressources. Ces inefficacités se traduisent directement par une augmentation des coûts logistiques, une dégradation du niveau de service et une complexification de la planification des tournées. Elles engendrent également des impacts environnementaux non négligeables, notamment en raison de la consommation énergétique accrue et des émissions associées aux collectes mal optimisées.

À grande échelle, les systèmes de logistique inverse impliquent des réseaux complexes et des flux hétérogènes, en particulier pour des résidus spécifiques tels que les huiles usagées, les lubrifiants et les bidons d'huile moteur. Ces flux, fortement liés aux activités de maintenance, présentent une variabilité importante et un caractère souvent intermittent, ce qui rend leur modélisation et leur prévision particulièrement complexes.

Dans ce contexte, la prévision des volumes de résidus à collecter constitue un levier stratégique pour améliorer la performance des systèmes de logistique inverse. Des prévisions plus fiables permettent d'optimiser la planification des tournées, de mieux allouer les ressources, de soutenir les décisions tactiques et stratégiques, et de concilier plus efficacement performance économique et objectifs environnementaux.

Ce projet de recherche a été réalisé dans le cadre du programme Mitacs Accélération, en collaboration avec deux partenaires industriels canadiens : RPM Eco et OSCAL AI.

RPM Eco est une entreprise canadienne spécialisée dans la collecte, la récupération et la valorisation de résidus industriels, notamment les contenants d'huile moteur usagés ainsi que d'autres résidus industriels. L'entreprise opère un vaste réseau de collecte à travers plusieurs régions du Canada, impliquant différents types de générateurs industriels et commerciaux. Les opérations de collecte sont caractérisées par une forte variabilité des volumes générés, des fréquences de collecte irrégulières ainsi qu'une dispersion géographique importante des sites.

OSCAL AI est une entreprise technologique spécialisée dans l'analyse de données et le déve-

loppement de solutions basées sur l’intelligence artificielle appliquées aux contextes industriels et logistiques. Dans le cadre de ce projet, OSCAL.AI contribue au développement d’approches analytiques et prédictives permettant de mieux exploiter les données opérationnelles issues des activités de collecte.

La collaboration entre RPM Eco et OSCAL AI vise à développer des outils d’aide à la décision basés sur les données afin d’améliorer la planification des opérations de collecte, l’allocation des ressources logistiques et l’efficacité globale des systèmes de récupération dans un contexte de logistique inverse à grande échelle.

Malgré l’importance opérationnelle de ces activités et la disponibilité croissante des données issues des opérations de collecte, plusieurs défis scientifiques et méthodologiques subsistent quant à la prévision fiable des flux de résidus industriels dans un contexte de logistique inverse.

1.2.1 Énoncé du problème et lacunes

Plusieurs limites subsistent dans les approches actuelles de prévision des résidus industriels.

Premièrement, les modèles statistiques traditionnels, tels que ARIMA et Croston, présentent des limites face à la complexité des flux industriels, notamment en raison de leur difficulté à capturer des relations non linéaires et des comportements fortement variables [5].

Deuxièmement, bien que les méthodes d’apprentissage automatique permettent d’améliorer la performance prédictive, elles exploitent encore de manière limitée la dimension spatiale et traitent généralement les unités géographiques de manière indépendante.

Troisièmement, les approches existantes intègrent rarement de manière conjointe les dimensions spatiale, temporelle et comportementale, limitant ainsi leur capacité à modéliser les interactions complexes entre unités géographiques.

Enfin, les mécanismes de dépendance spatiale, tels que les interactions entre zones voisines ou similaires, sont encore peu exploités dans les approches prédictives, malgré leur potentiel pour améliorer la robustesse des modèles.

Ces limites scientifiques motivent le développement d’approches intégrées combinant apprentissage automatique, structuration spatio-temporelle et mécanismes de régularisation spatiale.

Dans le contexte industriel étudié, les outils de prévision actuellement utilisés reposent principalement sur des approches statistiques traditionnelles et exploitent de manière limitée les données disponibles. En particulier, ces approches n’intègrent pas pleinement les dimen-

sions spatiales des collectes ni les interactions entre zones géographiques, et utilisent peu de variables contextuelles externes telles que les données météorologiques ou économiques.

Par ailleurs, les données opérationnelles sont généralement traitées de manière agrégée ou indépendante, sans tirer parti des structures spatiales sous-jacentes ni des dynamiques spatio-temporelles des flux de résidus.

Ces limites opérationnelles soulignent la nécessité de développer des approches plus avancées, capables d'intégrer simultanément plusieurs dimensions des données afin d'améliorer la précision et la robustesse des prévisions.

Ainsi, cette problématique scientifique et opérationnelle constitue le point de départ du présent mémoire.

1.2.2 Contributions à la recherche

Les principales contributions de ce mémoire sont les suivantes :

- Le développement d'un cadre de prévision spatio-temporel basé sur l'apprentissage automatique pour la génération de résidus industriels.
- L'intégration d'un système d'indexation spatiale hiérarchique (H3) afin de structurer les données géographiques et d'analyser l'impact de la granularité spatiale.
- L'évaluation comparative de plusieurs approches de modélisation, incluant des modèles statistiques traditionnels et des modèles d'apprentissage automatique (CatBoost, XGBoost, LightGBM).
- L'exploration de différentes stratégies d'agrégation temporelle (hebdomadaire, bihebdomadaire, mensuelle) et leur impact sur la stabilité et la performance des prédictions.
- L'étude de l'apport des variables météorologiques dans le processus de prévision, mettant en évidence leur contribution limitée dans le contexte étudié.
- La proposition d'un mécanisme de lissage spatial basé sur la similarité comportementale entre unités géographiques afin d'améliorer la robustesse des prévisions.
- Une analyse approfondie des compromis entre précision, granularité spatiale et stabilité des modèles dans un contexte de données industrielles réelles.

1.3 Plan de Mémoire

Le reste du mémoire est organisé comme suit. Chapitre 2 présente la revue de littérature, couvrant les principaux concepts liés à la logistique inverse, à la prévision de la demande intermittente, à l'apprentissage automatique et à la modélisation spatio-temporelle, ainsi que

les limites des approches existantes. Chapitre 3 présente le cadre méthodologique général adopté dans ce mémoire. Chapitres 4 et 5 regroupent les contributions scientifiques de ce mémoire sous forme de deux articles. Chapitre 4 présente le premier article, intitulé : *Optimizing Reverse Logistics : A Spatio-Temporal ML Framework for Industrial Waste Forecasting*, qui étudie l'effet de l'agrégation spatio-temporelle sur la performance prédictive. Chapitre 5 présente le deuxième article, intitulé : *Dynamic Spatial Smoothing Based on Production Profile Similarity for Industrial Waste Forecasting*, qui propose une amélioration des prévisions à fine résolution grâce à l'intégration de variables contextuelles et à un mécanisme de lissage spatio-temporel.

Enfin, Chapitre 6 conclut le mémoire en synthétisant les résultats, en discutant les implications de la recherche et en proposant des perspectives de travaux futurs.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Méthodologie de recherche documentaire

La revue de littérature a été réalisée principalement à partir de la base de données Compendex en utilisant des mots-clés liés à la logistique inverse, à la prévision des résidus industriels, à la demande intermittente et à l'apprentissage automatique.

Les travaux retenus ont été sélectionnés selon leur pertinence scientifique, leur contribution méthodologique et leur lien avec les problématiques étudiées dans ce mémoire, notamment la modélisation spatio-temporelle, l'agrégation géographique et les approches de prévision appliquées aux systèmes de logistique inverse.

Une attention particulière a été portée aux publications récentes parues entre 2015 et 2024 afin d'intégrer les avancées récentes dans le domaine.

La requête principale utilisée dans Compendex est la suivante :

((((waste NEAR/2 generation OR waste NEAR/2 quantit*) AND (model* OR algorithm* OR "machine learn*") AND (predict* OR forecast* OR estimat*) AND (reverse NEAR/2 logistic*)) wn KY

Le processus de sélection repose sur plusieurs étapes de filtrage et d'analyse des publications scientifiques pertinentes.

2.2 Logistique inverse et génération des résidus industriels

2.2.1 Logistique inverse et économie circulaire

La transition vers des modèles économiques plus durables a favorisé l'émergence de l'économie circulaire, qui vise à prolonger le cycle de vie des produits, à limiter l'extraction de nouvelles ressources et à réduire la génération des déchets par la réutilisation, le recyclage, le remanufacturing et la valorisation des matières [3], [4]. Contrairement au modèle linéaire traditionnel fondé sur la logique « extraire, produire, consommer, jeter », l'économie circulaire cherche à maintenir les ressources dans le système économique le plus longtemps possible.

Dans ce cadre, la logistique joue un rôle central puisqu'elle relie les activités de production, de consommation et de récupération des matières. La logistique directe (forward logistics) concerne l'acheminement des produits depuis les fournisseurs jusqu'aux utilisateurs finaux, alors que la logistique inverse (reverse logistics) porte sur les flux de retour de produits,

composants ou résidus depuis les points de consommation vers des centres de collecte, de traitement, de recyclage ou d'élimination sécuritaire [6], [7]. L'intégration de ces deux dimensions s'inscrit dans la logique des systèmes en boucle fermée (closed-loop systems), qui visent à coordonner les flux directs et inverses dans une perspective de circularité.

Toutefois, contrairement à la logistique directe, la logistique inverse se caractérise par une forte incertitude touchant la quantité, la qualité, la localisation et le moment d'apparition des retours [3]. Cette complexité est particulièrement marquée dans les systèmes industriels, où les flux dépendent à la fois du comportement des générateurs, de la durée de vie des équipements, des cycles d'activité et de facteurs externes [8], [9]. Les travaux récents montrent d'ailleurs que, malgré le développement d'outils numériques de suivi et de traçabilité, les systèmes de logistique inverse demeurent souvent pilotés de manière réactive plutôt que prédictive [10].

Dans les systèmes de gestion des résidus industriels, la logistique inverse ne se limite donc pas au simple transport des flux retournés. Elle implique des décisions stratégiques et opérationnelles liées à la planification des collectes, à l'allocation des ressources, au dimensionnement des capacités et à l'optimisation des réseaux de récupération. L'efficacité de ces décisions dépend fortement de la capacité à anticiper les volumes générés, ce qui place la prévision au cœur de la performance des systèmes inverses.

2.2.2 Variabilité et spécificité de la génération des résidus

La génération des résidus industriels présente des caractéristiques distinctes de la demande classique observée dans les chaînes d'approvisionnement directes. Elle est souvent intermittente, fortement variable, influencée par des facteurs contextuels et spatialement hétérogène. Cette variabilité se manifeste aussi bien dans le temps, à travers des alternances entre périodes creuses et pics de génération, que dans l'espace, selon les secteurs, les régions ou les profils de générateurs [11].

Les déchets industriels constituent en outre une catégorie très hétérogène. La littérature s'intéresse notamment aux déchets électroniques (*e-waste*), aux déchets huileux, aux véhicules en fin de vie (*end-of-life vehicles*) et aux déchets photovoltaïques, qui présentent chacun des contraintes particulières en matière de collecte, de traitement, de réglementation et de valorisation [12], [13]. Cette diversité rend difficile l'utilisation d'approches uniformes de planification et souligne la nécessité de développer des méthodes adaptées à la nature spécifique des flux étudiés.

Dans le contexte spécifique de cette étude, les résidus considérés correspondent aux bidons d'huile de moteur usagés, issus principalement des activités de maintenance automobile et

industrielle. Ces flux présentent des caractéristiques particulières, notamment une génération dépendante des cycles d'entretien, une forte variabilité selon les types de générateurs et une distribution spatiale hétérogène liée à l'activité des secteurs industriels et des ateliers mécaniques. Contrairement à d'autres types de déchets tels que les *e-waste*, ces résidus sont directement liés aux opérations de maintenance, ce qui induit des dynamiques spécifiques de génération.

Les modèles stochastiques développés pour la demande intermittente montrent que ce type de séries temporelles ne suit pas des dynamiques régulières et nécessite des approches spécifiques telles que la méthode de Croston [14]. Toutefois, ces méthodes reposent sur des hypothèses de stationnarité qui deviennent limitantes lorsque les volumes sont influencés par des facteurs exogènes, des comportements hétérogènes ou des conditions locales variables.

Dans le domaine des déchets municipaux et industriels, plusieurs études ont montré que les volumes collectés peuvent varier selon les saisons, les conditions météorologiques, l'activité économique ou les caractéristiques locales des générateurs [15], [13]. Cette variabilité rend la planification logistique particulièrement complexe, car une estimation inadéquate des volumes peut entraîner des inefficacités opérationnelles, une mauvaise allocation des ressources et des impacts environnementaux accrus.

2.2.3 Rôle de la prévision dans la performance des systèmes inverses

La performance des systèmes de logistique inverse dépend directement de la qualité des prévisions. En l'absence de modèles prédictifs robustes, les organisations s'appuient souvent sur des moyennes historiques ou des règles empiriques, ce qui limite leur capacité d'anticipation face à la variabilité réelle des flux.

La littérature sur la prévision de la demande souligne que l'intégration de méthodes quantitatives avancées permet d'améliorer significativement la précision et la robustesse des estimations [16], [17]. Toutefois, la majorité des travaux en prévision se concentre encore sur la demande directe, laissant un écart important dans l'application de ces approches aux flux inverses et aux résidus industriels.

Ainsi, malgré l'importance croissante de la logistique inverse dans les modèles circulaires, la prévision des résidus industriels demeure un champ de recherche nécessitant des approches adaptées aux caractéristiques intermittentes, spatiales et dynamiques des données. Ces constats mettent en évidence la nécessité d'identifier des méthodes de prévision capables de mieux représenter l'hétérogénéité des flux et de soutenir la prise de décision dans les systèmes de logistique inverse.

La section suivante présente les principales approches de prévision développées pour la demande intermittente ainsi que les avancées récentes en apprentissage automatique, en analysant leurs forces, leurs limites et leur pertinence dans un contexte industriel.

2.3 Prévision en logistique inverse et demande intermittente

2.3.1 Caractéristiques de la demande intermittente

La prévision des flux en logistique inverse présente des défis spécifiques liés à la nature intermittente et irrégulière des séries temporelles observées. Contrairement aux systèmes de distribution classiques, où la demande peut présenter des patterns relativement stables, les flux inverses sont souvent caractérisés par une alternance entre périodes nulles et occurrences ponctuelles de volumes positifs.

Les travaux fondateurs sur la demande intermittente ont montré que ces séries ne peuvent être modélisées efficacement par des méthodes traditionnelles telles que les modèles ARIMA standards, en raison de la présence fréquente de zéros et de la variabilité des intervalles entre deux occurrences successives. La méthode de Croston a été proposée pour traiter spécifiquement ce type de séries en séparant l'estimation de la taille de la demande et l'intervalle entre les occurrences. Les analyses stochastiques ultérieures ont toutefois mis en évidence certaines limites théoriques et biais potentiels associés à cette approche [18].

Dans les systèmes industriels, la génération de résidus peut présenter des comportements analogues à la demande intermittente, notamment lorsque certains clients génèrent des volumes uniquement à des moments spécifiques liés à leur cycle opérationnel.

2.3.2 Limites des approches classiques en contexte industriel

Malgré leur popularité, les méthodes classiques telles que Croston et ses variantes reposent sur des hypothèses de stationnarité qui peuvent être difficiles à satisfaire dans des environnements industriels dynamiques. De plus, ces méthodes considèrent généralement les séries individuellement, sans intégrer explicitement les dépendances spatiales ou les effets contextuels [5].

La littérature récente sur la prévision de la demande souligne que les modèles statistiques traditionnels montrent leurs limites lorsque les données présentent : une forte non-linéarité, des dépendances complexes, des influences exogènes multiples ou des interactions spatiales implicites [16]. Dans les systèmes de gestion des déchets, plusieurs études ont démontré que les volumes collectés sont influencés par des variables externes telles que les conditions météoro-

logiques ou les caractéristiques socio-économiques locales [15]. L'intégration de ces facteurs dépasse souvent les capacités des modèles purement linéaires.

Ces constats ont motivé le recours croissant aux méthodes d'apprentissage automatique pour la prévision des flux complexes.

2.3.3 Apport des méthodes d'apprentissage automatique

Les approches basées sur l'apprentissage automatique offrent une flexibilité accrue pour modéliser des relations non linéaires et capturer des interactions complexes entre variables [19]. Les modèles de gradient boosting, tels que XGBoost, LightGBM et CatBoost, ont démontré des performances supérieures dans de nombreux contextes de prévision, notamment en présence de données hétérogènes et de variables catégorielles [20].

Les revues récentes en prévision montrent que les modèles basés sur les réseaux neuronaux récurrents et les méthodes d'ensemble peuvent améliorer significativement la précision par rapport aux approches traditionnelles, en particulier lorsque les données proviennent de multiples séries temporelles interconnectées [17].

Cependant, malgré leurs performances élevées, les modèles d'apprentissage automatique appliqués aux données tabulaires présentent une limite importante : ils ne prennent généralement pas en compte de manière explicite la dimension spatiale des données. En l'absence d'une structuration géographique adaptée, ces modèles traitent les observations comme indépendantes, sans exploiter les relations de voisinage ou les corrélations spatiales potentielles.

Dans des systèmes de logistique inverse spatialement distribués, cette limitation peut conduire à une sous-exploitation de l'information disponible, en particulier lorsque des dépendances existent entre unités géographiques proches ou présentant des comportements similaires. Cette lacune constitue une motivation importante pour intégrer explicitement la dimension spatiale dans les modèles de prévision, comme discuté dans la section suivante.

À la lumière des limites identifiées dans les approches d'apprentissage automatique appliquées à des données tabulaires, l'intégration de la dimension spatiale apparaît comme une piste prometteuse pour améliorer la performance des modèles prédictifs.

En particulier, dans des contextes de logistique inverse, les unités géographiques ne sont pas indépendantes, mais peuvent présenter des interactions et des comportements similaires liés à leur proximité spatiale ou à des caractéristiques communes. Ne pas prendre en compte ces dépendances peut conduire à une perte d'information pertinente pour la prévision.

Dans cette perspective, des approches basées sur l'agrégation spatiale et le lissage spatial constituent des solutions adaptées pour structurer les données et exploiter les relations de

voisinage entre unités géographiques. Ces techniques permettent notamment de mieux capter les dynamiques locales et d'améliorer la robustesse des prédictions.

La section suivante présente ainsi les principaux travaux portant sur l'intégration de la dimension spatiale en prévision, en mettant l'accent sur les méthodes d'agrégation géographique et les approches de lissage spatial.

2.4 Dimension spatiale en prévision et agrégation géographique

2.4.1 Intégration de la dimension spatiale en prévision

Dans les systèmes de logistique inverse spatialement distribués, la génération des résidus ne dépend pas uniquement de dynamiques temporelles, mais également de la localisation géographique des sources. La littérature en économétrie spatiale souligne que les observations situées à proximité géographique tendent à présenter des comportements corrélés, un phénomène connu sous le nom d'autocorrélation spatiale [21].

Ignorer cette dépendance peut conduire à des estimations biaisées ou à une sous-exploitation de l'information disponible. Les approches classiques de prévision, même lorsqu'elles intègrent des variables exogènes, traitent généralement chaque série temporelle indépendamment, sans modéliser explicitement les interactions spatiales [22].

Dans les contextes urbains et industriels, plusieurs travaux ont montré que la prise en compte des corrélations spatiales améliore la performance des modèles prédictifs, notamment dans les applications de trafic, de mobilité et de gestion environnementale [23].

2.4.2 Problème d'agrégation spatiale et MAUP

L'un des enjeux majeurs en modélisation spatiale concerne le choix de l'unité géographique d'analyse. Le Modifiable Areal Unit Problem (MAUP) désigne la sensibilité des résultats statistiques au niveau d'agrégation spatial choisi [24]. Les estimations peuvent varier significativement selon que les données sont agrégées à une échelle fine ou à une échelle plus large.

Deux effets principaux sont associés au MAUP : l'effet d'échelle, qui correspond à la variation des résultats selon la taille des unités spatiales, et l'effet de zonage, qui reflète la variation liée à la définition des frontières géographiques [25].

Dans les systèmes de prévision, une agrégation excessive peut réduire la variance des données et améliorer artificiellement certains indicateurs statistiques, tels que le coefficient de détermination, tout en masquant la variabilité locale réelle. À l'inverse, une granularité très

fine peut accroître le bruit et rendre les prédictions plus instables [25].

Ce compromis entre résolution spatiale et prédictibilité constitue un enjeu central pour les systèmes de logistique inverse, où les décisions opérationnelles, telles que la planification des tournées ou l'allocation des ressources, sont prises à une échelle locale.

2.4.3 Indexation spatiale hiérarchique et système H3

Afin de structurer l'espace géographique de manière cohérente, plusieurs systèmes d'indexation spatiale hiérarchique ont été développés. Le système d'indexation spatiale hiérarchique hexagonale H3 (Hierarchical Hexagonal Spatial Indexing System) repose sur une discrétisation de la surface terrestre en cellules hexagonales organisées selon différents niveaux de résolution [26]. Cette structure permet une subdivision multi-échelle tout en conservant des propriétés géométriques régulières.

Comparativement aux grilles carrées traditionnelles ou aux découpages administratifs (par exemple les codes postaux), le maillage hexagonal présente plusieurs propriétés intéressantes pour l'analyse spatiale. Il permet notamment de réduire les biais directionnels et d'assurer une distance plus homogène entre le centre d'une cellule et ses voisines, ce qui facilite la modélisation des relations de proximité spatiale.

Ces caractéristiques peuvent être particulièrement pertinentes dans des contextes logistiques, où les interactions spatiales influencent directement la planification des tournées et l'organisation des collectes. De plus, le maillage hexagonal favorise une représentation plus régulière des relations d'adjacence et s'intègre bien dans des approches multi-résolution.

Cependant, l'utilisation de ce type de structuration spatiale présente également certaines limites. Tout d'abord, le choix du niveau de résolution reste un paramètre critique, car une granularité trop fine peut introduire du bruit et accentuer la variabilité des séries temporelles, tandis qu'une agrégation trop grossière peut masquer des dynamiques locales importantes. Par ailleurs, l'utilisation d'une grille artificielle, indépendamment des frontières administratives ou opérationnelles, peut entraîner une perte d'interprétabilité pour les décideurs.

Ainsi, bien que le maillage hexagonal constitue une approche pertinente pour structurer les données spatiales, son utilisation doit être soigneusement adaptée au contexte étudié, en tenant compte du compromis entre précision spatiale, stabilité des modèles et interprétabilité des résultats.

2.4.4 Implications pour la prévision des résidus industriels

Dans le contexte de la génération des résidus industriels, la distribution spatiale des clients induit une hétérogénéité géographique marquée. Une agrégation excessive peut masquer des dynamiques locales distinctes, tandis qu’une résolution trop fine accroît la variabilité et peut détériorer la stabilité prédictive. La littérature ne fournit pas de consensus clair quant au niveau d’agrégation optimal en prévision spatiale, et peu d’études combinent explicitement apprentissage automatique, indexation spatiale hiérarchique et analyse du compromis entre granularité et performance. Cette lacune justifie une évaluation systématique de l’impact de la résolution spatiale sur la précision des modèles, constituant ainsi le fondement méthodologique du premier article de ce mémoire.

Ainsi, bien que les approches d’apprentissage automatique améliorent significativement la précision des prévisions, l’intégration explicite de la dimension spatiale et la détermination d’un niveau d’agrégation géographique approprié demeurent des enjeux méthodologiques ouverts. L’impact de la résolution spatiale sur la performance des modèles prédictifs nécessite donc une analyse systématique.

2.5 Lissage spatio-temporel et approches par diffusion

2.5.1 Dépendance spatiale et lissage

Lorsque les données sont modélisées à une résolution spatiale fine, la variance locale augmente significativement. Bien que cette granularité permette de capturer des comportements spécifiques, elle peut également introduire un bruit important et réduire la stabilité des prévisions.

La littérature en économétrie spatiale souligne que les observations géographiquement proches présentent souvent des structures corrélées [21]. Cette dépendance spatiale suggère que l’information contenue dans les unités voisines peut être exploitée afin d’améliorer les estimations locales.

Dans les modèles spatio-temporels appliqués au trafic, à la mobilité ou aux systèmes urbains, l’intégration des dépendances spatiales améliore significativement la robustesse des prédictions [22, 23]. Ces approches reposent sur l’idée que les dynamiques locales ne sont pas indépendantes, mais s’inscrivent dans une structure spatiale plus large.

Ainsi, lorsque la modélisation est réalisée à une résolution fine (comme dans le cas d’une indexation hexagonale H3), la question du lissage spatial devient centrale afin de réduire le bruit tout en conservant l’information locale pertinente.

2.5.2 Modèles de diffusion sur graphes

Les modèles de diffusion spatio-temporelle formalisent cette intuition en considérant les unités spatiales comme des nœuds d'un graphe, reliés par des relations de voisinage. L'évolution temporelle d'un nœud est alors influencée non seulement par son historique propre, mais également par les signaux observés dans ses voisins.

Les travaux récents en traitement du signal sur graphes montrent que la régularisation par diffusion permet de récupérer des signaux bruités en exploitant une hypothèse de lissage spatial [16]. Ces approches supposent que les signaux évoluent de manière relativement cohérente au sein d'un voisinage géographique.

Plus récemment, les modèles de diffusion spatio-temporelle appliqués aux graphes dynamiques ont démontré des performances élevées en prévision multi-séries lorsque les dépendances spatiales sont explicitement intégrées dans l'architecture du modèle [17].

Cependant, la majorité de ces travaux repose sur des architectures neuronales complexes, nécessitant des volumes de données importants et une infrastructure computationnelle élevée. Leur applicabilité dans des contextes industriels, où l'interprétabilité et la robustesse sont essentielles, peut être limitée.

2.5.3 Similarité comportementale et voisinage adaptatif

Au-delà du voisinage purement géographique, plusieurs travaux en prévision spatio-temporelle ont montré que la similarité comportementale entre unités spatiales peut constituer une base pertinente pour la régularisation [27].

Le voisinage géographique repose sur la proximité physique entre unités (i.e., les zones adjacentes), tandis que le voisinage comportemental ou fonctionnel repose sur la similarité des dynamiques de génération observées dans le temps. Autrement dit, deux unités sont considérées comme voisines non pas parce qu'elles sont proches spatialement, mais parce qu'elles présentent des profils de génération similaires.

Plutôt que de considérer uniquement la proximité physique, ces approches exploitent la corrélation des profils temporels afin de définir des voisinages fonctionnels. Cette logique est également présente dans les méthodes de clustering de séries temporelles, où des groupes sont formés sur la base de comportements similaires [28].

Par exemple, deux zones industrielles géographiquement éloignées peuvent présenter des dynamiques de génération de résidus plus similaires entre elles qu'avec une zone résidentielle adjacente, en raison de la nature comparable de leurs activités économiques. À l'inverse, deux

unités spatialement proches peuvent présenter des comportements très différents.

Dans les systèmes de génération de résidus industriels, intégrer cette dimension comportementale permet ainsi de dépasser les limites d'un lissage strictement spatial et de mieux capturer la structure réelle des données.

2.5.4 Limites actuelles et opportunités de recherche

Malgré les avancées en modélisation spatio-temporelle, plusieurs lacunes subsistent dans la littérature. Peu d'études combinent explicitement une indexation spatiale hiérarchique, telle que le système H3, avec des mécanismes de lissage adaptatif. Les approches par diffusion demeurent majoritairement appliquées aux domaines du trafic ou de l'énergie et sont rarement mobilisées pour des flux de logistique inverse. Par ailleurs, l'effet d'un lissage pondéré par similarité comportementale dans un cadre d'apprentissage automatique tabulaire reste peu exploré. Ces constats suggèrent qu'une approche intégrant une granularité spatiale fine, une modélisation par apprentissage automatique et une régularisation fondée sur la similarité des profils et la structure de voisinage pourrait renforcer la stabilité et la performance prédictive des systèmes de gestion des résidus industriels.

2.6 Clustering spatio-temporel et regroupement dynamique

2.6.1 Clustering de séries temporelles : fondements théoriques

Le clustering vise à regrouper des entités présentant des comportements similaires afin de révéler des structures latentes dans les données. Dans le contexte des séries temporelles, l'objectif est d'identifier des groupes partageant des dynamiques communes en termes de tendance, de saisonnalité ou de variabilité. Les travaux comparatifs récents en clustering de séries temporelles montrent que la performance dépend fortement de la représentation choisie (features statistiques vs séries brutes), de la mesure de distance utilisée et de la stabilité temporelle des groupes [18].

Dans les applications prédictives, le clustering peut être utilisé comme étape préalable ("cluster-then-forecast") afin d'améliorer la cohérence des modèles en regroupant des unités homogènes.

2.6.2 Clustering spatial et segmentation géographique

Dans les systèmes géographiquement distribués, le clustering peut également intégrer une dimension spatiale. Plusieurs études en segmentation de marché et en géo-analyse ont démontré que le regroupement d'unités partageant des caractéristiques communes améliore la

robustesse des décisions stratégiques [19].

Toutefois, le clustering purement spatial peut être insuffisant lorsque la similarité comportementale ne correspond pas strictement à la proximité géographique. Dans le contexte des résidus industriels, deux hexagones adjacents peuvent présenter des volumes très différents selon l'activité des clients présents.

Cela souligne la nécessité d'une approche hybride intégrant la structure spatiale, la similarité comportementale et la dynamique temporelle.

2.6.3 Clustering dynamique et adaptation temporelle

Les systèmes industriels évoluent dans le temps. Les comportements de génération de résidus peuvent changer en raison des modifications opérationnelles, d'évolutions économiques, de variations saisonnières ou d'introduction de nouvelles politiques environnementales.

Un clustering statique peut devenir obsolète si les profils évoluent. La littérature en segmentation dynamique montre que des regroupements recalculés périodiquement permettent d'améliorer la stabilité des modèles prédictifs dans des environnements non stationnaires [18].

Cependant, peu d'études appliquent explicitement le clustering dynamique à la prévision des flux de logistique inverse, encore moins dans un cadre combinant indexation spatiale hiérarchique et apprentissage automatique.

2.7 Synthèse de la revue de la littérature

Afin de synthétiser les principaux travaux recensés dans la littérature et de mettre en évidence leurs contributions ainsi que leurs limites, le Tableau 2.1 présente une comparaison des études les plus pertinentes.

TABLE 2.1 Synthèse de la revue de littérature

Auteur	Approche	Contexte	Étude de cas	Critères considérés
Lieder et al. (2016)	Conceptuelle	Économie circulaire	Industrie manufacturière	Stratégiques
Acerbi et al. (2020)	Conceptuelle	Logistique durable	Industrie	Stratégiques
Hyndman et al. (2008)	Statistique	Demande intermittente	Séries temporelles	Temporels
Fildes et al. (2019)	Statistique	Demand forecasting	Multi-secteurs	Temporels
Bandara et al. (2020)	Machine learning	Prévision multi-séries	Données synthétiques / réelles	Temporels
Souza et al. (2022)	Hybride (ARIMA–ANN)	Déchets industriels	Véhicules en fin de vie	Temporels
Ducharme et al.	Machine learning	Gestion des déchets	Données municipales	Temporels et environnementaux
Lee et Shin (2025)	Machine learning (RF, XGBoost)	Déchets urbains	Données socio-économiques	Temporels et contextuels
Li et al. (2015)	Spatio-temporel	Mobilité / trafic	Véhicules urbains	Spatiaux et temporels
Wang et al. (2020)	Spatio-temporel	Prévision multi-zones	Données urbaines	Spatiaux
Soares et al. (2025)	Indexation spatiale (H3)	Modélisation géospatiale	Trajectoires aériennes	Spatiaux
Chen et al.	Machine learning (boosting)	Données tabulaires	Applications diverses	Temporels
Javed et al.	Clustering	Séries temporelles	Benchmark	Temporels
Kimiagari et al.	Clustering	Segmentation	Marketing / spatial	Spatiaux
Ce mémoire	ML + spatio-temporel + lissage	Logistique inverse	Résidus industriels (Canada)	Temporels, spatiaux, comportementaux

La revue de la littérature met en évidence que la prévision des flux en logistique inverse constitue un domaine en évolution, caractérisé par une forte variabilité temporelle, une hétérogénéité spatiale marquée et une dépendance à des facteurs contextuels.

Les approches traditionnelles, notamment celles développées pour la demande intermittente, offrent un cadre analytique de base, mais présentent des limites lorsqu'elles sont appliquées à des environnements industriels complexes et spatialement distribués.

Les méthodes d'apprentissage automatique permettent de mieux capturer les relations non linéaires et d'intégrer un grand nombre de variables explicatives, améliorant ainsi la performance prédictive. Toutefois, ces modèles considèrent généralement les unités spatiales de manière indépendante et exploitent peu la structure géographique sous-jacente.

Par ailleurs, la littérature souligne l'importance du choix de l'unité spatiale d'analyse, notamment à travers la problématique du Modifiable Areal Unit Problem (MAUP), qui influence directement la stabilité et la précision des résultats.

Dans l'ensemble, ces travaux mettent en évidence la nécessité de développer des approches intégrées capables de combiner les dimensions spatiale, temporelle et comportementale afin d'améliorer la précision et la robustesse des modèles de prévision.

À la lumière des limites identifiées dans la littérature, ce mémoire propose deux contributions complémentaires visant à améliorer la modélisation de la génération des résidus industriels dans un contexte de logistique inverse :

(i) l'étude de l'impact de l'agrégation spatio-temporelle sur la performance des modèles de prévision, afin de mieux comprendre le compromis entre résolution des données et stabilité des prédictions, et (ii) le développement d'une approche de lissage spatial fondée sur la similarité comportementale entre unités géographiques, dans le but d'exploiter les dépendances spatiales souvent négligées dans les approches classiques.

Ces contributions s'inscrivent dans une démarche visant à intégrer de manière conjointe les dimensions spatiale, temporelle et comportementale, afin de mieux capturer la complexité des flux de résidus industriels.

2.8 Lacunes et limites des approches existantes

L'analyse de la littérature permet d'identifier plusieurs limites majeures dans les approches actuelles de prévision des résidus industriels :

- **Gap 1 – Absence d'analyse de la granularité spatiale :** La majorité des travaux reposent sur des niveaux d'agrégation spatiale fixes, sans étudier systématiquement

l’impact de la granularité spatiale sur la performance des modèles prédictifs.

- **Gap 2 – Indépendance des unités spatiales dans les modèles d’apprentissage automatique :** Bien que les méthodes d’apprentissage automatique améliorent significativement la précision des prévisions, elles traitent généralement les unités spatiales de manière indépendante et n’intègrent pas explicitement les interactions spatiales.
- **Gap 3 – Limites des approches de lissage spatial classiques :** Les méthodes de lissage ou de diffusion spatiale existantes reposent principalement sur la proximité géographique et exploitent peu la similarité comportementale entre unités spatiales.
- **Gap 4 – Faible intégration des variables contextuelles :** L’intégration de variables contextuelles, telles que les conditions météorologiques, reste encore peu étudiée dans le contexte spécifique de la logistique inverse industrielle.

Ces lacunes mettent en évidence l’opportunité de développer des approches intégrées combinant granularité spatiale adaptée, modélisation robuste par apprentissage automatique et prise en compte des interactions spatiales et comportementales.

2.9 Positionnement scientifique du mémoire

La revue de littérature met en évidence que la prévision des flux en logistique inverse demeure un domaine en évolution, caractérisé par une forte variabilité temporelle, une hétérogénéité spatiale marquée et une dépendance potentielle aux facteurs contextuels. Bien que les méthodes traditionnelles, telles que celles développées pour la demande intermittente, offrent un cadre analytique initial, elles présentent des limites importantes lorsqu’elles sont appliquées à des environnements industriels complexes et spatialement distribués.

Les approches d’apprentissage automatique permettent de modéliser des relations non linéaires et d’intégrer des variables explicatives multiples, améliorant ainsi la performance prédictive. Toutefois, ces modèles traitent généralement les unités spatiales de manière indépendante et n’intègrent pas explicitement la structure géographique sous-jacente aux données. Par ailleurs, le choix du niveau d’agrégation spatiale influence significativement la stabilité et la précision des estimations, comme le souligne la problématique du Modifiable Areal Unit Problem (MAUP).

Malgré l’existence de systèmes d’indexation spatiale hiérarchique tels que H3, leur impact sur la performance des modèles de prévision en logistique inverse reste encore peu étudié de manière systématique. De plus, les approches de régularisation spatio-temporelle, notamment celles basées sur la similarité comportementale entre unités géographiques, demeurent largement sous-explorées dans ce contexte.

Dans ce cadre, ce mémoire vise à combler ces lacunes en proposant une approche intégrée combinant apprentissage automatique, structuration spatio-temporelle des données et mécanismes de lissage spatial. Plus précisément, le premier article analyse de manière systématique l'effet de la granularité spatiale définie par le système H3 sur la performance prédictive. Le second article introduit un mécanisme de lissage spatio-temporel fondé sur la similarité comportementale afin d'améliorer la robustesse des prévisions à haute résolution spatiale.

CHAPITRE 3 CADRE MÉTHODOLOGIQUE GÉNÉRAL

3.1 Objectifs de recherche

L'objectif principal de ce mémoire est de développer un cadre prédictif spatio-temporel permettant d'améliorer la prévision de la génération des résidus industriels dans un contexte de logistique inverse, en s'appuyant sur des techniques d'apprentissage automatique et sur une structuration adaptée des données.

Afin d'atteindre cet objectif, ce travail s'articule autour de deux objectifs spécifiques :

1. **Analyser l'impact de l'agrégation spatio-temporelle sur la performance prédictive** : Cet objectif vise à étudier l'influence du choix de la granularité spatiale et temporelle, notamment à travers l'utilisation du système d'indexation hexagonale H3, sur la qualité des prévisions des résidus industriels.
2. **Développer une approche de lissage spatio-temporel basée sur la similarité comportementale** : Cet objectif consiste à améliorer la précision des prévisions à fine résolution spatiale en proposant un mécanisme de régularisation exploitant la similarité des profils de génération entre unités géographiques, tout en intégrant des variables explicatives pertinentes telles que les conditions météorologiques.

Ce mémoire propose ainsi une approche intégrée permettant de mieux capturer la complexité spatio-temporelle des flux de résidus industriels et d'améliorer la planification des opérations de collecte. Afin de répondre à ces objectifs de recherche, une démarche méthodologique structurée a été développée, combinant des approches d'apprentissage automatique, une structuration spatio-temporelle des données ainsi qu'un mécanisme de régularisation spatiale.

3.2 Méthodologie générale

La démarche méthodologique générale adoptée dans ce mémoire repose sur plusieurs étapes complémentaires, incluant la préparation et l'intégration des données, la structuration spatio-temporelle des observations, le développement des modèles prédictifs, ainsi que l'évaluation des performances à différentes échelles spatiales et temporelles.

L'approche proposée combine des techniques d'apprentissage automatique avec des mécanismes de structuration spatiale afin de mieux capturer la variabilité et l'intermittence des flux de résidus industriels.

Ce mémoire est structuré sous la forme de deux articles scientifiques complémentaires. La Figure 3.1 illustre le cadre général de la démarche de recherche ainsi que l’articulation des contributions proposées.

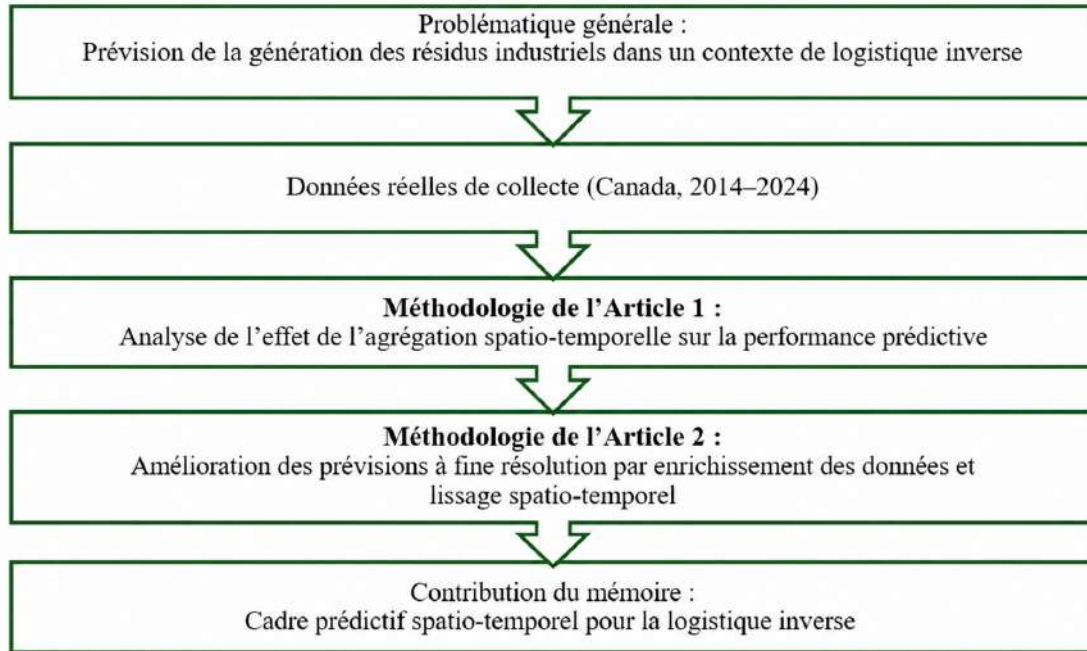


FIGURE 3.1 Cadre conceptuel du projet de recherche et articulation des deux articles.

3.3 Structure des articles

3.3.1 Article 1 : Prévision de la génération des résidus industriels

Le premier article propose un cadre de prévision basé sur l'apprentissage automatique pour estimer la génération des résidus industriels. Les données sont structurées à l'aide du système d'indexation spatiale H3 et agrégées sur des périodes bi-hebdomadaires afin de capturer les dynamiques spatio-temporelles. Plusieurs modèles sont ensuite évalués afin d'identifier l'approche la plus performante, en s'appuyant notamment sur une validation croisée temporelle (TimeSeriesSplit) ainsi que sur une procédure systématique de recherche d'hyperparamètres basée sur une stratégie de type random search, permettant d'explorer efficacement l'espace des configurations possibles. La Figure 3.2 illustre les principales étapes méthodologiques du premier article, incluant le prétraitement des données, la construction des variables explicatives, l'entraînement des modèles et leur évaluation.

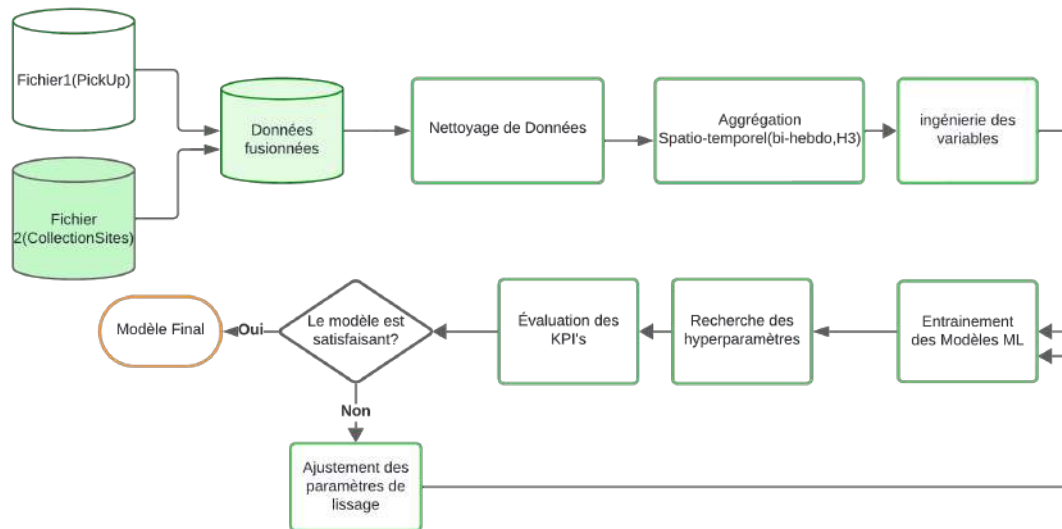


FIGURE 3.2 Pipeline méthodologique de l'Article 1 : structuration spatio-temporelle et modélisation prédictive.

Ce schéma met en évidence la structuration progressive des données ainsi que l'évaluation des modèles prédictifs dans un cadre spatio-temporel.

3.3.2 Article 2 : Amélioration des prévisions par lissage spatio-temporel

Le deuxième article s'inscrit dans la continuité directe du premier. Alors que l'Article 1 met en évidence l'importance des stratégies d'agrégation spatio-temporelle et permet d'obtenir des performances prédictives élevées à l'aide de modèles d'apprentissage automatique, certaines limites subsistent, notamment en présence de forte variabilité locale et de données peu denses à haute résolution spatiale. Dans un premier temps, des variables contextuelles, notamment les conditions météorologiques, ont été explorées afin d'enrichir les modèles prédictifs. Toutefois, leur contribution à l'amélioration des performances s'est révélée limitée dans le contexte étudié. Afin de répondre à ces limitations, l'Article 2 propose une extension du cadre initial en introduisant un mécanisme de lissage spatio-temporel fondé sur la similarité comportementale entre unités géographiques. Cette approche permet d'exploiter les dépendances spatiales et d'améliorer la robustesse des prévisions, en particulier dans les zones caractérisées par des séries temporelles instables ou intermittentes. La Figure 3.3 explique les principales étapes méthodologiques du deuxième article.

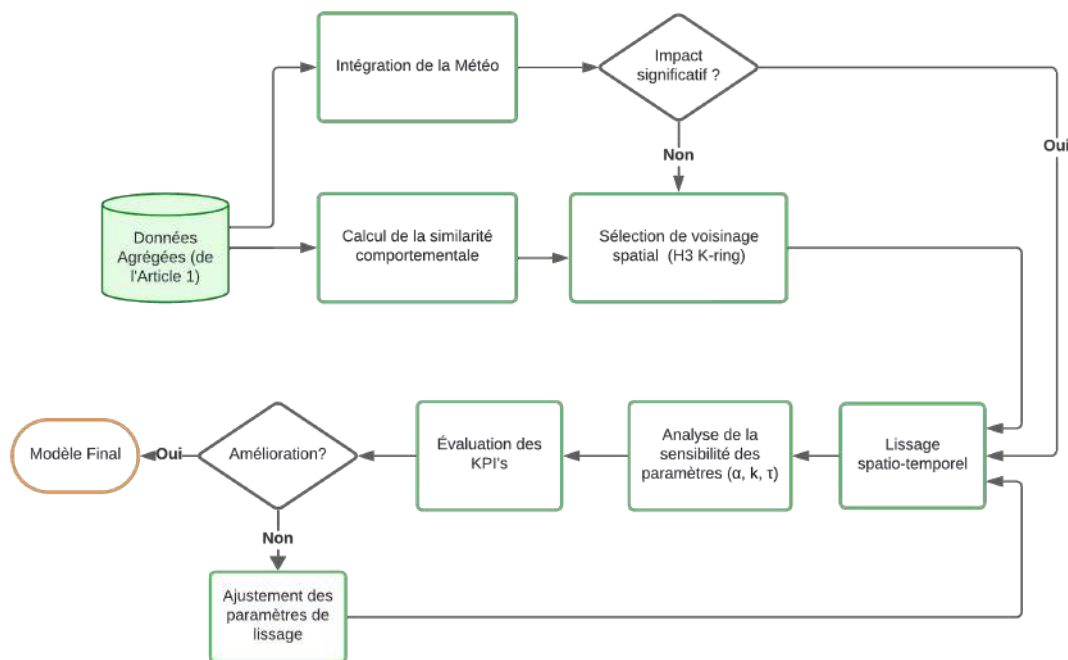


FIGURE 3.3 Approche proposée dans l'Article 2 : lissage spatio-temporel basé sur la similarité.

3.3.3 Préparation et enrichissement des données

Avant le développement des modèles prédictifs, plusieurs étapes de préparation et d'enrichissement des données (*feature engineering*) ont été réalisées afin de mieux représenter les dynamiques observées dans les flux de résidus industriels.

Ces étapes incluent notamment le nettoyage des données, le traitement des valeurs aberrantes ainsi que la construction de variables explicatives adaptées au contexte étudié.

Des variables temporelles ont été extraites à partir des dates de collecte, telles que le mois, le trimestre et la semaine de l'année, afin de capturer les effets de saisonnalité. Des variables de retard (*lags*) ont également été générées (par exemple $t - 1$, $t - 2$, $t - 3$), permettant de modéliser les dépendances temporelles.

Par ailleurs, des statistiques glissantes ont été calculées sur différentes fenêtres temporelles (3 et 5 périodes), incluant la moyenne et l'écart-type, afin de capturer les tendances locales et la variabilité des flux.

Étant donné la nature temporelle des données étudiées, une validation croisée aléatoire classique n'a pas été utilisée afin d'éviter toute fuite d'information (*data leakage*) entre les ensembles d'entraînement et de test. En effet, dans un contexte de séries temporelles, un mé-

lange aléatoire des observations pourrait conduire à l'utilisation d'informations futures pour prédire des observations passées, produisant ainsi une estimation artificiellement optimiste des performances prédictives.

L'évaluation des modèles repose donc principalement sur une séparation chronologique des données, où les périodes les plus anciennes sont utilisées pour l'entraînement tandis que les périodes plus récentes servent à la validation et au test. Cette stratégie permet de reproduire plus fidèlement le contexte opérationnel réel, dans lequel les prédictions sont toujours effectuées à partir d'informations historiques disponibles au moment de la décision.

Par ailleurs, certaines phases d'ajustement des hyperparamètres et d'expérimentation ont été réalisées à l'aide d'approches de validation adaptées aux séries temporelles, notamment des stratégies de type *TimeSeriesSplit*. Cette approche repose sur une fenêtre glissante (*rolling window*) dans laquelle l'ensemble d'entraînement est progressivement élargi au fil du temps, tandis que les périodes suivantes sont utilisées pour la validation. Ainsi, l'ordre temporel des observations est systématiquement préservé tout au long du processus d'évaluation.

Cette stratégie de validation permet d'obtenir une estimation plus robuste et plus réaliste des performances des modèles dans un contexte de prévision opérationnelle des résidus industriels.

Enfin, des indicateurs comportementaux ont été construits au niveau des unités de collecte, tels que la moyenne historique, la fréquence des collectes et des ratios par rapport aux valeurs historiques, permettant de caractériser les profils de génération des résidus.

3.4 Justification des choix méthodologiques

3.4.1 Choix des modèles d'apprentissage automatique

Les modèles d'apprentissage automatique ont été privilégiés dans ce mémoire en raison de leur capacité à capturer des relations non linéaires complexes ainsi que des dynamiques spatio-temporelles difficilement modélisables par les approches statistiques traditionnelles telles que ARIMA ou Croston.

Parmi les approches évaluées, les modèles basés sur le gradient boosting, notamment CatBoost, XGBoost et LightGBM, ont été retenus en raison de leurs performances reconnues sur les données tabulaires, de leur robustesse face à l'hétérogénéité des données et de leur capacité à gérer efficacement des ensembles de données de grande dimension. Par ailleurs, les modèles tabulaires basés sur le gradient boosting présentent également l'avantage d'offrir un meilleur compromis entre performance prédictive, robustesse et interprétabilité comparativement à certaines approches de deep learning plus complexes.

3.4.2 Justification du système d'indexation H3

Le système d'indexation spatiale hiérarchique H3 a été utilisé afin de structurer les données spatiales à l'aide d'un maillage hexagonal homogène. Cette approche permet de mieux représenter les variations spatiales des flux de résidus industriels tout en réduisant certains biais liés aux découpages administratifs traditionnels.

L'utilisation d'une structuration hexagonale facilite également l'analyse des relations de voisinage spatial ainsi que l'intégration de mécanismes de lissage spatial basés sur les interactions entre unités géographiques.

3.4.3 Choix de l'agrégation temporelle

Différentes granularités temporelles ont été explorées dans les phases initiales de cette recherche, notamment des agrégations hebdomadaires, bi-hebdomadaires et mensuelles. L'agrégation bi-hebdomadaire a finalement été retenue puisqu'elle offrait le meilleur compromis entre stabilité des séries temporelles et performance prédictive. Cette agrégation permet de réduire l'effet du bruit et des fluctuations très court terme tout en conservant une résolution temporelle adaptée aux besoins opérationnels de planification.

3.4.4 Mécanisme du lissage spatio-temporel

Enfin, un mécanisme de lissage spatio-temporel a été proposé afin d'améliorer la robustesse des prédictions à fine résolution spatiale. Cette approche repose sur l'exploitation de la similarité comportementale entre unités géographiques ainsi que sur la prise en compte du voisinage spatial à l'aide de l'opérateur k-ring du système H3. La similarité entre unités spatiales est mesurée à l'aide de la distance euclidienne, choisie pour sa simplicité, sa stabilité numérique et sa capacité à capturer efficacement les différences globales entre profils temporels normalisés. Bien que d'autres métriques puissent être envisagées, cette mesure offre un bon compromis entre performance et interprétabilité dans le cadre étudié. Plusieurs configurations ont été testées, incluant différentes tailles de voisinage, paramètres de pondération et mesures de similarité.

3.5 Présentation du cas d'étude et des données

Ce mémoire s'appuie sur un cas d'étude industriel réel portant sur la collecte de bidons d'huile de moteur usagés au Canada dans un contexte de logistique inverse.

Le réseau étudié comprend différents types de générateurs, notamment des garages automo-

biles, ateliers mécaniques, centres de maintenance ainsi que certains sites industriels répartis sur plusieurs régions canadiennes. Les fréquences de collecte varient selon les profils des clients, les volumes générés ainsi que les contraintes opérationnelles propres à chaque site. Certains générateurs présentent des comportements relativement stables alors que d'autres produisent des flux intermittents et fortement variables, compliquant la planification des opérations de collecte.

Les données utilisées proviennent de notre partenaire industriel RPM Eco spécialisé dans la récupération et la valorisation des résidus industriels. Elles couvrent une période d'observation allant de 2014 à 2024 et comprennent plus d'un million d'observations issues des opérations réelles de collecte.

Deux principales bases de données ont été exploitées dans le cadre de cette recherche.

La première base de données, intitulée *PickUp-PlasticContainer-Oil*, contient les informations transactionnelles liées aux opérations de collecte. Cette base comprend environ 1 765 057 observations et 26 variables décrivant les collectes réalisées. Les principales variables incluent notamment les identifiants des sites de collecte (*CollectionSiteID*), les dates planifiées et réalisées des collectes (*PlannedDate*, *RealisedDate*), les quantités récupérées (*WeightPickUp*), ainsi que différentes informations opérationnelles telles que les seuils minimaux et maximaux de collecte (*ActualMin*, *ActualMax*), le type de collecte (*TypePickUp*) et certains indicateurs logistiques (*DSLP*). La Figure 3.4 présente un extrait de la base de données *PickUp-PlasticContainer-Oil* utilisée pour représenter les opérations de collecte réalisées sur le terrain.

TaskID	CollectionSite	RouteID	PlannedDate	RealisedDate	WeightPickUp	WeightPredict	SectorID	Statut	TypePickUp	DSLP	ServiceRepCa	CommentsClt	ObservedLat	ObservedLong	ActualMin	ActualMax
AB2149-004-29	CC028182	AB2149-004	2021-11-29	2021-12-03	0	70	817 Max-CT4	ForceService		179 NULL			51.2885869	-113.9803066	39	65
AB2149-005-1	CC026014	AB2149-005	2021-11-30	2021-11-30	0	120	619 Max	Prediction		17 NULL			53.17535	-113.6131126	52	104
AB2149-006-1	CC021649	AB2149-006	2021-11-30	2021-12-02	100	254	823 Max	Prediction		12 NULL			53.7911636	-113.5905899	117	221
AB2149-006-2	CC021804	AB2149-006	2021-11-30	2021-12-02	3	39	823 Min	Prediction		219 NULL			53.6334758	-113.623321	39	65
AB2149-006-3	CC017784	AB2149-006	2021-11-30	2021-12-02	70	71	823 Min	Prediction		6 NULL			53.646381	-113.6278234	35	85
AB2149-006-4	CC028390	AB2149-006	2021-11-30	2021-12-02	84	70	823 Max-CT2	ForceService		27 NULL			53.6186417	-113.6430417	65	120
AB2149-006-5	CC021537	AB2149-006	2021-11-30	NULL	0	53	823 Max-CT8	ForceService		408 NULL	NULL	NULL	NULL	NULL	39	65
AB2149-006-6	CC021536	AB2149-006	2021-11-30	2021-12-02	84	24	823 Max-CT4	ForceService		28 NULL			53.646381	-113.6278234	39	65
AB2149-006-7	CC016675	AB2149-006	2021-11-30	2021-12-02	14	29	823 Min	Prediction		201 NULL			53.6188657	-113.6438021	26	52
AB2149-006-8	CC015722	AB2149-006	2021-11-30	2021-12-02	14	35	823 Min	Prediction		306 NULL			53.6334887	-113.623336	20	39
AB2149-006-9	CC019861	AB2149-006	2021-11-30	2021-12-02	14	41	823 Min	Prediction		299 NULL			53.5402343	-113.7959923	39	65
AB2149-006-10	CC026106	AB2149-006	2021-11-30	2021-12-02	58	35	823 Min	Prediction		119 NULL			53.5683006	-113.7662456	20	39
AB2149-006-11	CC019660	AB2149-006	2021-11-30	2021-12-02	14	35	823 Min	Prediction		168 NULL			53.5402343	-113.7959923	26	52
AB2149-006-12	CC017637	AB2149-006	2021-11-30	2021-12-02	0	49	823 Min	Prediction		380 NULL			53.5419008	-113.8864426	26	52
AB2149-006-13	CC014146	AB2149-006	2021-11-30	2021-12-02	28	42	823 Min	Prediction		120 NULL			53.566485	-113.7582683	39	65
AB2149-010-26	CC026378	AB2149-010	2021-12-01	2021-11-30	0	69	821 Max	Prediction		7 NULL			53.17535	-113.6131126	20	39
AB2149-007-5	CC026975	AB2150-008	2021-12-01	2021-12-08 27.36		39	836 Max-CT1	ForceService		8 NULL			55.2284867	-118.61671	40	100
AB2149-007-6	CC019224	AB2149-007	2021-12-01	NULL	NULL	58	836 Min	Prediction		240 NULL	NULL	NULL	NULL	NULL	52	104
AB2149-007-8	CC028060	AB2150-008	2021-12-01	2021-12-08 60.9		34	836 Max-CT3	ForceService		36 NULL			55.2284867	-118.61671	23	59
AB2149-007-9	CC021578	AB2149-007	2021-12-01	NULL	NULL	16	836 OPP	Prediction		197 NULL	NULL	NULL	NULL	NULL	20	39
AB2149-007-11	CC020224	AB2149-007	2021-12-01	NULL	NULL	68	836 Min	Prediction		91 NULL	NULL	NULL	NULL	NULL	52	104
AB2149-007-12	CC027261	AB2149-007	2021-12-01	NULL	NULL	39	836 OPP-CT13	Prediction		92 NULL	NULL	NULL	NULL	NULL	20	39

FIGURE 3.4 Extrait de la base de données *PickUp-PlasticContainer-Oil*.

La seconde base de données, nommée *CollectionSites-PlasticContainer-Oil*, contient les in-

formations descriptives associées aux sites de collecte. Elle comprend environ 28 952 observations et 33 variables. Cette base fournit notamment les coordonnées géographiques des générateurs (*Latitude*, *Longitude*), les villes (*City*), ainsi que diverses informations administratives et opérationnelles relatives aux clients. La Figure 3.5 présente un extrait de la base de données *CollectionSites-PlasticContainer-Oil*, contenant les informations descriptives et géographiques associées aux différents sites de collecte.

CollectionSiteID	CollectionSite Address	ZipCode	City	Latitude	Longitude	ZoneID	Province	PhoneNumbe	SectorID	Minimum	Maximum	DistanceFrom	ClientTypeID	CollectionCer	PredictionID	AddedDate	Active
CC33511	Drolet Ressor 1275 Rue de l' G1N 1X8		Quebec	46.8006889	-71.2579113	Z2QC		4186875222	217	39		65 267.954535	CT2	CCTR3	PT1	2023-10-27	TRUE
CC33570	HDM Diesel St 3153 NS-395 B0E 3M0		East Lake Ains	46.1216833	-61.1424715	ZANS		9029562194	127	39		65 3743.902088	CT4	CCTR4	PT1	2023-11-23	TRUE
CC33523	Kelly Auto Rep 1335 Humber K1B 5K9		Ottawa	45.4024622	-75.6091787	Z2ON		6137412626	386	39		65 171.930542	CT4	CCTR3	PT1	2023-11-02	FALSE
CC33524	Kal Tire 42 Isnor Dr B3B 1W3		Dartmouth	44.7062955	-63.5982239	Z1NS		9024688473	131	39		65 250.614172	CT4	CCTR4	PT1	2023-11-03	TRUE
CC33527	GFL Environm 444 Advance / K7R 3Z6		Napanee	44.2608581	-76.9578014	Z2ON		6133549393	255	39		65 0	CT4	CCTR3	PT1	2023-11-06	FALSE
CC33531	Lucid Motor 790 Terminal / V6A 2M5		Vancouver	49.2696536	-123.0864447	Z1BC		7786534666	905	39		65 27.397412	CT100	CCTR5	PT1	2023-11-07	FALSE
CC33532	Bishops Cran 52 Terminal R A1X 7B6		Conception Bt	47.512226	-52.9781574	Z1NL		7098342080	101	39		65 15.634489	CT31	ACNL4	PT1	2023-11-08	TRUE
CC33591	Sheppard Lan 12111 68 St St T2Z 0E5		Calgary	50.9445828	-113.9371485	Z4AB		4032684904	833	39		65 334.121318	CT4	CCTR1	PT1	2023-11-24	FALSE
CC33536	Desjardins Au 250 Rue Etien G1M 3K6		Quebec	46.8296778	-71.2730666	Z2QC		0	217	39		65 273.254707	CT3	CCTR3	PT1	2023-11-09	TRUE
CC33537	Trail Tire West 1708 37 Stree T3C 3R1		Calgary	51.0386834	-114.139977	Z4AB		4032491048	832	39		65 324.030549	CT8	CCTR1	PT1	2023-11-09	TRUE
CC33540	Trail Tire - Falc McInight villag T3J 3H4		Calgary	51.0982155	-113.9551137	Z4AB		4032933001	829	39		65 313.005413	CT12	CCTR1	PT1	2023-11-09	TRUE
CC33541	Trail Tire - 17 / 3420 17 Ave S T2A 0R4		Calgary	51.0384672	-113.9860077	Z4AB		4032722266	833	39		65 319.0477	CT16	CCTR1	PT1	2023-11-09	TRUE
CC33542	Wiman paving 220 Edworthy V3L 5G6		New Westmin	49.223113	-122.8873838	Z1BC		6045223921	2008	45		65 17.615982	CT5	CCTR5	PT1	2023-11-09	TRUE
CC33543	Trail Tire - McIn 9250 Macleod T2J 0P5		Calgary	50.9705965	-114.0714795	Z4AB		4032538004	832	39		65 329.778016	CT4	CCTR1	PT1	2023-11-09	TRUE
CC33546	Hamilton Trac 1425 Hwy 56 N3W 1T1		Caledonia	43.0661192	-79.8107984	Z2ON		2894420701	518	39		65 38.043473	CT5	CCTR2	PT1	2023-11-10	TRUE
CC33549	That Auto Rep 443 Exeter Rd N6E 2Z3		London	42.9234851	-81.2370205	Z2ON		5196867253	512	39		65 137.226524	CT10	CCTR2	PT1	2023-11-13	TRUE
CC33551	Crewe Motors 776 Dundas St L4Y 2B6		Mississauga	43.5960437	-79.5976757	Z1ON		4163122645	506	39		65 34.617798	CT6	CCTR2	PT1	2023-11-14	TRUE
CC33555	J. Martin Truc 11423 NB-12E E4Y 2N7		Rogersville	46.7519572	-65.4486736	Z3NB		5066253702	148	39		65 97.002229	CT20	CCTR4	PT1	2023-11-16	TRUE
CC33602	Mr. Lube 2505 Eversyde T2Y 5G8		Calgary	50.90796	-114.1177	Z4AB		8255401338	832	39		65 341.784332	CT4	CCTR1	PT1	2023-12-01	TRUE
CC33512	Pacific Coast 44250 Progre V2R 3W8		Chilliwack	49.1540258	-121.999618	Z1BC		6047749305	922	45		65 87.378116	CT4	CCTR5	PT1	2023-10-27	FALSE
CC33515	Yacht Club St 1350 Chem. d H9S 2E3		Donval	45.4390944	-73.7538187	Z1QC		5146312720	376	39		65 38.504629	CT100	CCTR3	PT1	2023-10-30	TRUE

FIGURE 3.5 Extrait de la base de données *CollectionSites-PlasticContainer-Oil*.

Les deux bases de données ont été fusionnées à l'aide de l'identifiant unique *CollectionSiteID* afin d'enrichir les observations de collecte par leurs caractéristiques spatiales et contextuelles. Cette intégration a permis de construire une représentation spatio-temporelle détaillée des flux de résidus industriels à travers le territoire canadien.

Les données présentent une forte variabilité spatio-temporelle ainsi qu'un caractère souvent intermittent. Certaines zones géographiques génèrent des volumes relativement stables alors que d'autres présentent des comportements très irréguliers ou fortement fluctuants. Cette hétérogénéité rend la modélisation prédictive particulièrement complexe et justifie l'utilisation d'approches avancées d'apprentissage automatique combinées à des mécanismes de structuration spatiale et de régularisation.

Ce chapitre a présenté le cadre méthodologique général adopté dans le cadre de ce mémoire, incluant les objectifs de recherche, la démarche globale de modélisation ainsi que les principaux choix méthodologiques liés à la structuration spatio-temporelle des données et au développement des modèles prédictifs.

Le cas d'étude industriel ainsi que les données utilisées ont également été présentés afin de contextualiser les approches développées dans ce travail.

Le chapitre suivant présente le premier article scientifique de ce mémoire, qui porte sur

l'analyse de l'impact des stratégies d'agrégation spatio-temporelle sur la performance des modèles de prévision des résidus industriels.

CHAPITRE 4 ARTICLE 1 : OPTIMIZING REVERSE LOGISTICS : A SPATIO-TEMPORAL ML FRAMEWORK FOR INDUSTRIAL WASTE FORECASTING

Submitted on March 10, 2026 and Published on August 21-23, 2026 in ICTLE 2026 14th
International Conference on Traffic and Logistics Engineering
Kso, A., Keivanpour, S., Frayret, J (2026, March 10)

Abstract : Accurate forecasting of industrial waste generation is critical for efficient reverse logistics operations, yet remains challenging due to temporal variability, spatial heterogeneity, and intermittent demand patterns. Most existing studies rely on fixed or highly aggregated spatial units, limiting their applicability for short- and medium-term operational planning. This paper proposes a machine learning-based forecasting framework that integrates spatio-temporal aggregation strategies to enhance predictive accuracy and operational relevance. Temporal aggregation is employed to reduce demand intermittence, while spatial aggregation is investigated at multiple levels, including city-based aggregation and fine-grained hexagonal partitioning using the H3 spatial indexing system at resolution 6. Using large-scale, real-world data collected in Canada, this study investigates the prediction of industrial waste through machine learning models, including CatBoost, XGBoost, and LightGBM, with classical statistical methods (ARIMA and Croston) serving as baselines. Experimental results demonstrate that gradient boosting models consistently outperform traditional statistical methods, achieving error reductions of up to 30% in terms of RMSE and MAE. The results further reveal that the choice of spatial aggregation granularity has a significant impact on forecast accuracy, stability, and interpretability. Overall, the proposed framework provides scalable and localized waste generation forecasts, supporting more informed operational planning and resource allocation in industrial reverse logistics systems.

Keywords : Reverse logistics, Industrial waste forecasting, Spatio-temporal aggregation, Machine learning, Spatial clustering, Demand forecasting.

4.1 Introduction

Reverse logistics is a pillar of circular economy systems, facilitating the collection, recovery, and reintegration of industrial waste into value chains. In contrast to forward logistics,

reverse logistics operations are subject to high uncertainty regarding waste generation quantities, collection frequency, and spatial distribution, which complicates planning and resource allocation. Industrial waste generation is typically intermittent and highly variable across locations, making accurate forecasting a persistent operational challenge [29]. Traditional forecasting approaches such as deterministic models and classical time series methods like ARIMA have achieved moderate success for aggregated and stationary waste flows [9]. However, their performance declines in complex industrial environments where data exhibit non-linear behaviors, irregular temporal patterns, and frequent zero-demand periods. Recent advances in machine learning (ML) offer new opportunities to capture these dynamics by modeling non-linear relationships among multiple explanatory variables [30]. Nevertheless, most existing ML-based studies apply fixed spatial units and rarely examine how spatial and temporal aggregation choices affect forecasting accuracy and stability. In operational practice, the level of spatial and temporal aggregation strongly influences predictive performance. Coarse administrative boundaries may obscure local variability, whereas overly detailed spatial partitions can amplify intermittence and noise. Similarly, temporal aggregation can enhance stability but risks compromising responsiveness if not optimally calibrated. To address these limitations, this paper proposes an ML-based framework for industrial waste generation forecasting that combines temporal aggregation and multi-scale spatial aggregation. Forecasting performance is assessed under city-level and hexagonal spatial partitioning using the H3 indexing framework, combined with a biweekly temporal aggregation scheme designed to balance intermittence reduction and operational relevance. Classical statistical models are benchmarked against advanced machine learning algorithms using real-world industrial data. The study investigates how aggregation affects forecasting results.

4.2 Literature Review

This section reviews the existing literature related to industrial waste generation forecasting within reverse logistics systems. It first outlines the fundamental characteristics and challenges of reverse logistics operations, then surveys existing approaches to waste generation forecasting. Next, it examines recent advances in machine learning-based prediction models, together with spatial and temporal aggregation strategies proposed to handle demand variability and heterogeneity. Finally, it identifies the main research gaps in order to clearly position the contribution of this study.

4.2.1 Reverse Logistics and Industrial Waste Generation

Reverse logistics is widely recognized as a key enabler of circular economy systems, as it governs the collection, treatment, and recovery of post-consumption and industrial waste flows. Recent literature emphasizes that reverse logistics systems are inherently characterized by high uncertainty in waste quantities, timing, and spatial distribution, which significantly complicates planning and operational decision-making [29], [30]. Industrial waste streams such as oily waste, electronic waste, end-of-life vehicles, and photovoltaic waste present additional challenges due to regulatory constraints, heterogeneous generation patterns, and high collection and treatment costs [31], [32]. Recent systematic reviews and bibliometric analyses highlight that, despite the maturity of conceptual frameworks for reverse and closed-loop logistics, their operational implementation remains limited by the lack of reliable forecasting tools capable of anticipating waste generation at decision-relevant spatial and temporal scales [12].

4.2.2 Waste Generation Forecasting in Reverse Logistics

Early waste generation forecasting approaches relied mainly on deterministic scenarios, life-cycle distribution models, and macroeconomic indicators, particularly for long-term planning of photovoltaic waste and end-of-life vehicles [33]. Although these methods are useful for strategic infrastructure planning, they lack the flexibility required for operational decision making in reverse logistics. Time series forecasting models such as ARIMA and SARIMA have been widely applied to waste forecasting problems and have demonstrated acceptable performance for stable and aggregated waste streams [34]. However, these models are inherently univariate and assume stationarity, which limits their applicability in industrial reverse logistics contexts characterized by non-linear dynamics, high variability, and heterogeneous generation behaviors [35].

4.2.3 Machine Learning and Advanced Forecasting Approaches

Recent studies highlight the strong potential of ML techniques for improving waste generation forecasting by capturing complex and non-linear relationships between multiple influencing variables. Supervised ML algorithms such as artificial neural networks, random forests, gradient boosting, and ensemble models have demonstrated superior predictive accuracy compared to traditional statistical methods, particularly in multivariate industrial contexts [36]. Hybrid forecasting approaches, combining multiple ML models or integrating statistical and ML-based predictors, have shown promising results in industrial and energy-related applica-

tions, including petroleum product traffic forecasting and waste-related flows [37]. However, most existing studies focus on aggregated datasets and do not fully exploit fine-grained operational data at the collection site level. Furthermore, recent systematic and bibliometric reviews confirm that ML applications in reverse logistics have predominantly focused on optimization and descriptive analytics, while operational predictive modeling remains underexplored [12], [38].

4.2.4 Geographical Clustering and Operational Gaps

The accounting of spatial heterogeneity is a critical challenge in reverse logistics systems, particularly when waste generation patterns vary significantly between locations. Geographical clustering techniques have therefore been widely used to group waste generators or regions with similar spatial characteristics, mainly for network design, market segmentation, or strategic profiling purposes [38]. Recent studies show that traditional geographical clustering approaches based on administrative boundaries or Euclidean distance often fail to capture the spatial continuity and local heterogeneity required for operational decision-making. To address these limitations, hexagonal hierarchical spatial indexing systems, such as the H3 framework, have emerged as promising alternatives [39]. H3 enables the discretization of geographical space into uniform hexagonal cells with multi-resolution properties, allowing consistent spatial aggregation and comparison across spatial scales. Recent applications in transportation and urban analytics have demonstrated that H3-based spatial frameworks are effective in capturing localized spatial patterns while maintaining scalability for large-scale operational analyzes [40]. Despite its increasing adoption in transportation, mobility, and urban service analytics, the use of H3-based spatial clustering remains largely unexplored in the context of industrial reverse logistics and waste generation forecasting. Existing studies that consider spatial clustering typically treat it as an exploratory or standalone step, disconnected from predictive modeling tasks [31].

Existing studies have explored the integration of spatial information into forecasting models, including approaches such as spatio-temporal autoregressive models and Bayesian hierarchical frameworks. However, these methods are often developed in different application contexts and remain relatively complex to implement in industrial reverse logistics settings.

Moreover, spatial clustering and segmentation are typically treated as exploratory or standalone steps, rather than being directly integrated into predictive modeling pipelines. As a result, their potential to improve forecasting accuracy or to support operational decision-making, such as route planning, resource allocation, or service differentiation, remains underexploited in practical applications. The gap between spatial clustering and forecasting model

integration limits the practical applicability of current analytical approaches in reverse logistics. In particular, there is a lack of hierarchical spatial clustering frameworks capable of identifying localized waste generation patterns while remaining scalable and interpretable at the operational decision-making level.

4.2.5 Research Gap and Positioning of This Paper

The literature review reveals two major gaps. First, most waste generation forecasting studies rely on fixed spatial and temporal aggregation levels, without systematically analyzing how aggregation choices influence forecasting stability and accuracy. This limits the applicability of existing models for short and medium term operational planning in reverse logistics systems. Second, in the context of industrial waste generation forecasting in reverse logistics systems, temporal, spatial, and contextual variables are rarely integrated within unified predictive frameworks.

To address these gaps, this article proposes an ML-based framework to forecast industrial waste generation in a reverse logistics context. The proposed approach integrates multivariate forecasting with hierarchical geographical clustering based on spatial H3 indexing to capture heterogeneity among waste generators and support operational decision-making. Through the use of fine-grained industrial data, this study aims to bridge the gap between advanced predictive analytics and practical reverse logistics applications.

4.3 STRUCTURE AND CONTEXT

4.3.1 Industrial Context and Partners

This study is carried out in collaboration with an industrial partner, RPM Eco, a Canadian company specializing in the collection and recovery of used motor oil containers within a reverse logistics framework. The firm operates a large-scale collection network that serves industrial and commercial generators across multiple urban and semi-urban regions in Canada. The project also involves OSCAL AI, a startup focused on data analytics and artificial intelligence solutions for environmental and logistics applications. The collaboration aims to design data-driven tools to enhance operational planning, with a specific emphasis on collection scheduling, resource allocation, and routing optimization. The industrial setting is marked by high uncertainty in waste generation, complex collection frequencies, and a heterogeneous spatial distribution of generators, which makes accurate forecasting both critical and challenging.

4.3.2 Datasets Description

This study adopts two main datasets provided by the industrial partner, which are related to the management of empty motor oil plastic containers. The first data set, called Pickup-PlasticContainer-Oil, contains historical records of collection operations and comprises 1,765,057 observations and 26 variables from 2014 to 2024. Each record corresponds to an actual collection event and includes, among others, a unique collection task identifier, the associated collection site identifier, the realized collection date, the collected weight of empty oil containers, and several operational or service-related attributes. The collected weight of used oil containers is the target variable of this study, as it represents the quantity of waste generated and collected at each service event and thus constitutes the demand signal to be forecast in the reverse logistics system. The second dataset, CollectionSites-PlasticContainer-Oil, describes the waste generators (collection sites) and contains 28,952 records and 33 attributes, providing spatial and contextual information such as geographical coordinates (latitude and longitude), city information, sector identifiers, and service-related characteristics. This study merges the collection sites dataset with the pickup dataset using the unique site identifier, enabling spatial analyses and the construction of aggregated demand profiles. Such large-scale industrial datasets are rarely available in academic studies, which strengthens the practical relevance of this research.

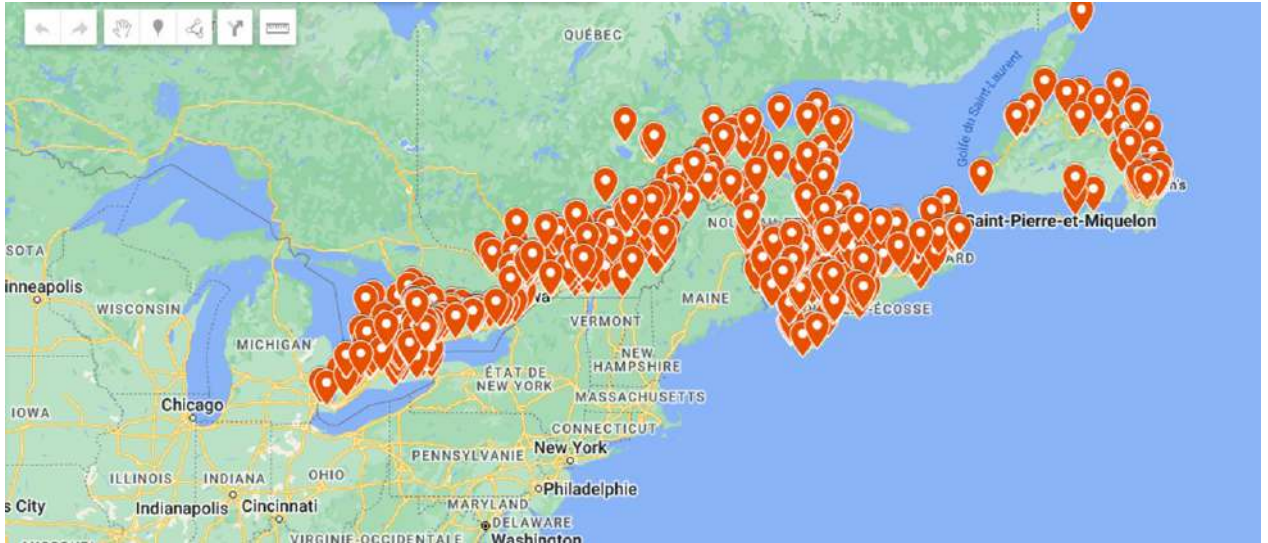


FIGURE 4.1 Spatial distribution of industrial collection sites across Eastern Canada.

4.3.3 Data Characteristics and Challenges

The industrial data exhibit several characteristics that make forecasting particularly challenging. First, waste generation is highly irregular over time. Collection events do not occur at fixed intervals, leading to intermittent time series with a large proportion of zero-demand periods when analyzed at fine temporal resolutions. Second, the data present strong spatial heterogeneity. Waste generation patterns differ significantly across locations, reflecting variations in customer activity levels, service policies, and local operational constraints. Third, the dataset is large-scale, exceeding one million records, which requires scalable modeling approaches capable of handling both temporal depth and spatial diversity. These characteristics motivate the need for appropriate spatio-temporal aggregation strategies to reduce noise, manage intermittence, and extract stable demand patterns suitable for operational forecasting.

4.3.4 Study Area

The study focuses on the geographical regions covered by the industrial partner’s collection network, where collection sites are located in several Canadian cities with a strong concentration in major metropolitan areas. To investigate spatial effects at different levels of detail, two spatial representations are considered :

- Administrative aggregation at the city level, commonly used in operational reporting,
- Fine-grained spatial aggregation using hexagonal cells, based on the H3 hierarchical spatial indexing system.

This dual spatial representation makes it possible to systematically compare traditional administrative aggregation with grid-based spatial discretization for waste generation forecasting.

4.4 Methodology

4.4.1 Overview of the Proposed Framework

The proposed methodology aims to assess how spatio-temporal aggregation strategies affect industrial waste generation forecasting performance. It relies on a structured pipeline that transforms raw collection data into predictive inputs and evaluates multiple forecasting models under various aggregation settings. Concretely, the approach consists of sequential steps : data cleaning and preprocessing, temporal aggregation of pickup events, spatial aggregation at different resolution levels, feature engineering, and model training and evaluation. This structured design supports consistent comparison across aggregation strategies and forecas-

ting methods, while ensuring both reproducibility and operational relevance. As illustrated in Fig. 4.2, the proposed framework integrates data preprocessing, spatio-temporal aggregation, feature engineering, model training, and performance evaluation within an iterative learning process.

4.4.2 Data cleaning and preprocessing

The raw datasets provided by the industrial partner contain several years of operational records and, as is common in real reverse logistics settings, include inconsistencies, missing values, and heterogeneous formats that must be corrected before modeling. Therefore, a dedicated data cleaning and preprocessing pipeline was designed to ensure data quality and consistency before aggregation and forecasting.

First, collection records with missing or invalid values in key fields collection date, collected weight, or site identifier were discarded to avoid bias in aggregated demand series and preserve temporal continuity. Date variables were then standardized into a unified datetime format : timestamps were checked for chronological consistency, and erroneous or duplicate entries were removed, enabling reliable temporal aggregation and preventing artificial peaks or gaps in the demand time series. Collected weights were subsequently screened to detect abnormal or non-physical values. Extremely huge or negative quantities, as well as implausible outliers likely due to measurement or reporting errors, were filtered out using domain informed thresholds, thereby stabilizing the statistical properties of the forecasting targets, especially after temporal aggregation.

To enrich the data with spatial information, the collection dataset was joined with the collection-sites database via a unique site identifier, associating each pickup event with attributes such as latitude, longitude, and city. Records with missing or inconsistent geographic coordinates were corrected when possible or removed to maintain spatial coherence. Geographic coordinates were then validated and consistently projected so that each site could be assigned to its corresponding spatial unit, either a city or a hexagonal cell generated with the H3 spatial indexing system.

Overall, this preprocessing pipeline transforms raw industrial records into a clean, structured, and spatially coherent dataset, suitable for subsequent spatio-temporal aggregation, feature engineering, and forecasting model development.

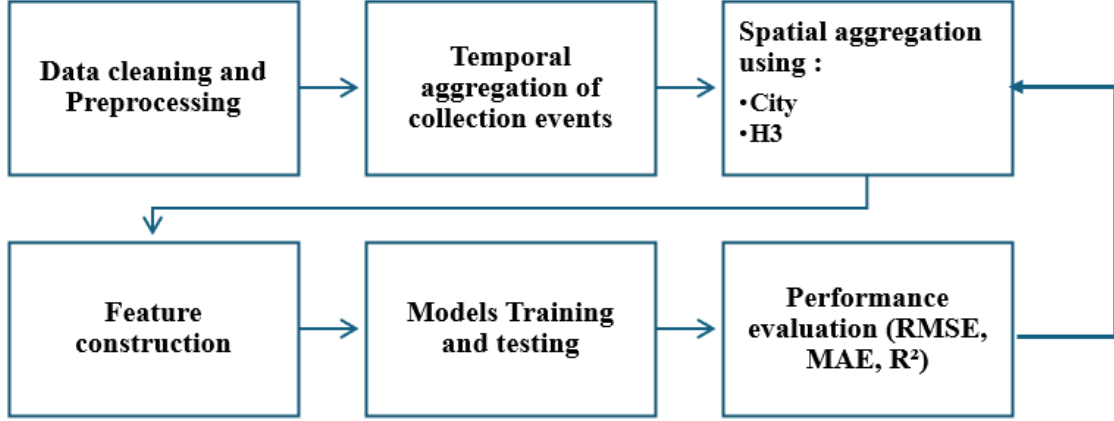


FIGURE 4.2 Overall methodological workflow adopted in this study.

4.4.3 Temporal Aggregation Strategy

At the raw data level, collection events occur at irregular dates and form highly intermittent signals [40], which results in very sparse time series with many zero-demand periods and unstable statistical properties [41], [42]. To address this, a biweekly (2-week) temporal aggregation is applied, where all collection events within the same two-week window are summarized by the total collected weight, the number of pickups, and the number of active collection sites. This aggregation horizon is chosen as a compromise between reducing intermittence and noise while preserving sufficient temporal resolution for operational decision-making : compared with weekly aggregation, it substantially decreases sparsity, yet remains more responsive than monthly aggregation for short- and medium-term planning. Fig. 4.3 highlights the strong irregularity of weekly demand patterns, which motivates the adoption of a bi-weekly temporal aggregation strategy.

4.4.4 Spatial Aggregation Strategies

To analyze spatial effects, two spatial aggregation levels are considered.

City-Level Aggregation In the first approach, collection data are aggregated by city and time period, reflecting the level of aggregation typically used by industrial operators for reporting and planning purposes. This aggregation level provides a baseline representation of waste generation patterns at a coarse spatial scale.

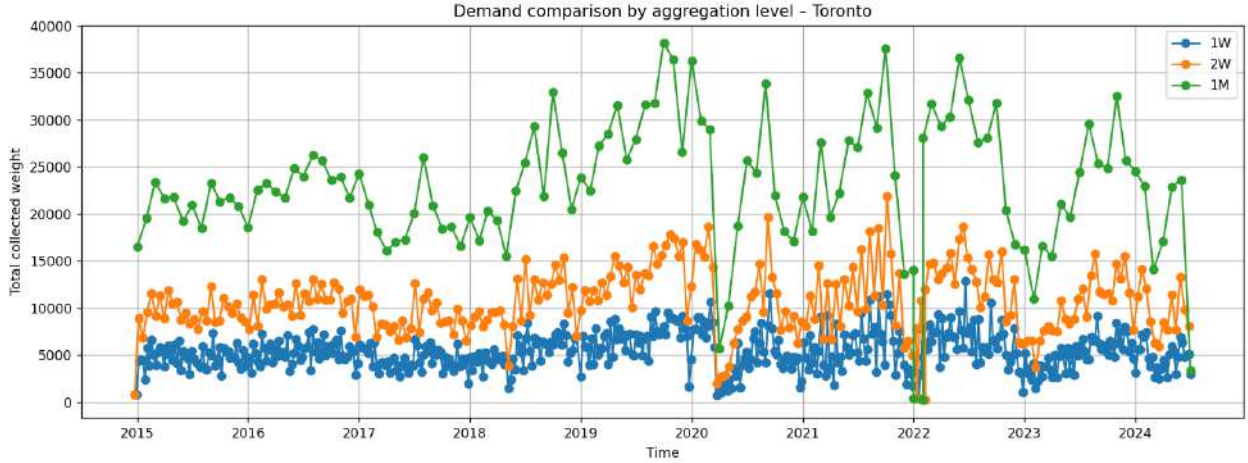


FIGURE 4.3 Comparison of waste generation patterns under different temporal aggregation levels (weekly, biweekly, and monthly) for the Toronto area

H3-Based Hexagonal Aggregation To better capture fine-grained spatial heterogeneity, a second aggregation scheme based on the H3 hierarchical hexagonal spatial indexing system is used [39]. Each collection site is assigned to an H3 cell using its geographical coordinates, and aggregation is performed at the (H3 cell, biweekly) level. This representation provides uniform hexagonal cells, is independent of administrative boundaries, offers a hierarchical multi-resolution structure, and scales well to large spatial datasets. Multiple H3 resolutions are evaluated to assess how spatial granularity affects forecasting performance. As illustrated in Fig. 4.4, the H3 indexing system enables a uniform hexagonal spatial partitioning of the collection sites, independent of administrative boundaries.

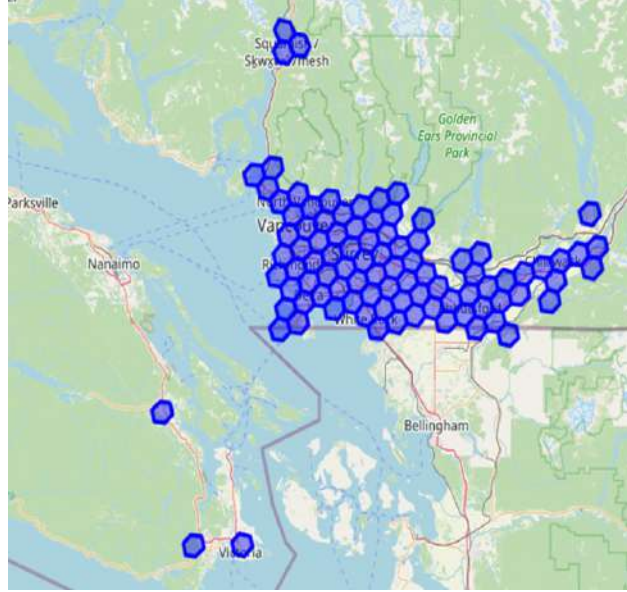


FIGURE 4.4 Hexagonal spatial aggregation of collection sites using the H3 indexing system.

4.4.5 Feature Construction

For each spatial unit (city or H3 cell), time series features are constructed to capture temporal dynamics and recent demand history. The complete set of engineered features used for waste generation forecasting is summarized in Table 4.1. These features are grouped into temporal, statistical, calendar-based, operational, and spatial categories in order to capture both short-term variations and structural demand patterns. Calendar-based features such as month, quarter, week index, and season were not originally available in the raw dataset and were derived from collection dates to enrich the temporal representation of waste generation patterns. Rolling standard deviation features capture the volatility of waste generation over recent periods, enabling machine learning models to account not only for average demand levels but also for fluctuations and uncertainty in waste generation patterns. These volatility indicators are particularly important in intermittent demand contexts, where sudden changes in generation behavior frequently occur.

TABLE 4.1 Summary of engineered features used for waste generation forecasting

Category	Feature	Description
Lagged demand features	y_{t-1}	Collected weight during the previous time period
	y_{t-2}	Collected weight two periods before
	y_{t-3}	Collected weight three periods before
Rolling statistics	Rolling mean (3)	Mean collected weight over the last three periods
	Rolling mean (5)	Mean collected weight over the last five periods
	Rolling std (3)	Standard deviation over the last three periods
	Rolling std (5)	Standard deviation over the last five periods
Calendar-based features	Month	Month of the year
	Quarter	Quarter of the year
	Week index	Index of the period within the year
	Season	Categorical variable (winter, spring, summer, fall)
Operational intensity features	Number of pickups	Total number of collection events within the period
	Number of active sites	Number of distinct collection sites served during the period
Spatial identifier	City / H3 cell ID	Identifier of the spatial aggregation unit

These features enable machine learning models to learn both short-term temporal dependencies and recurring seasonal patterns.

4.4.6 Forecasting Models

Baseline Models Two baseline forecasting methods commonly used in industrial settings are considered : ARIMA, which is well suited to aggregated and relatively stable time series, and Croston, which is specifically designed for intermittent demand. These models are used as benchmarks to evaluate the added value of the proposed models.

ML Models Three gradient boosting-based ML models are evaluated : CatBoost, XGBoost, and LightGBM. These models are particularly strong on structured tabular data and accurately learn non-linear relationships among explanatory variables. CatBoost handles categorical features and reduces overfitting through ordered boosting, XGBoost offers robust regularization and scalability, while LightGBM provides high computational efficiency via histogram-based learning [38]. These models are selected due to their strong performance on tabular data, ability to model non-linear relationships, and robustness to heterogeneous features [43]. Hyperparameter tuning is performed using optuna, and the final models are trained using the optimal parameter configurations.

4.4.7 Training and Evaluation Protocol

To preserve the temporal structure of the data, a time-based splitting strategy was adopted. The dataset was divided chronologically, where historical observations were used for training and more recent periods were reserved for validation and testing. This approach prevents information leakage from future observations into past periods and better reflects real-world forecasting conditions.

In addition to the main chronological train-test split, some hyperparameter tuning and experimental analyses were conducted using time-series cross-validation strategies based on rolling or expanding windows (*TimeSeriesSplit*). This validation procedure preserves the temporal ordering of the observations while providing a more robust evaluation of model stability across multiple forecasting periods.

Model performance was assessed using three standard metrics : Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). All models were evaluated using identical temporal splits to ensure fair comparability across spatial aggregation schemes and forecasting methods.

4.5 Results and Discussions

This section presents the experimental results obtained under different spatial and temporal aggregation schemes. The objective is to evaluate (i) the impact of spatial aggregation, (ii) the performance of ML models compared to baselines, and (iii) the contribution of spatio-temporal aggregation to forecasting accuracy.

4.5.1 City-Level Forecasting Results

Table 4.2 presents the forecasting performance of the evaluated models at the city level using biweekly temporal aggregation.

TABLE 4.2 Performance comparison of forecasting models (City, biweekly)

Model	RMSE (kg)	MAE (kg)	R ²
ARIMA	528.51	173.65	0.7127
Croston	514.95	169.38	0.7272
LightGBM	414.42	124.85	0.8135
XGBoost	403.38	121.65	0.8233
CatBoost	397.72	125.03	0.8283

ML models clearly outperform classical statistical baselines. While ARIMA and Croston capture general temporal trends, their linear structure and univariate formulation limit predictive accuracy. Gradient boosting models substantially improve forecasting performance, with CatBoost achieving approximately 25% RMSE reduction compared to ARIMA, demonstrating its ability to model non-linear temporal dynamics following temporal aggregation. The high R² values superior than 0.82 confirm that city-level aggregation generates stable demand patterns well suited for predictive modeling, making this spatial scale particularly effective for medium-term operational planning.

4.5.2 H3-Based Spatial Forecasting Results

Table 4.3 presents the forecasting performance obtained using hexagonal spatial aggregation at H3 resolution 6.

At the H3 level, the absolute error values decrease due to the lower magnitude of the demand per spatial unit. However, the overall predictive performance measured by the determination coefficient (R^2) deteriorates compared to the aggregation at the city-level. This behavior is

TABLE 4.3 Performance comparison of forecasting models (H3, biweekly)

Model	RMSE (kg)	MAE (kg)	R^2
ARIMA	557.32	255.01	0.2633
Croston	328.53	126.99	0.3866
LightGBM	267.10	90.30	0.5614
XGBoost	264.17	88.21	0.5710
CatBoost	259.30	88.94	0.5867

primarily explained by the intrinsic characteristics of fine grained spatial segmentation, which introduces increased spatial sparsity, higher intermittency, and shorter historical time series within individual hexagonal cells. As a result, a larger proportion of variability remains unexplained, leading to lower R^2 values despite reduced absolute errors. Nevertheless, ML models consistently outperform statistical baselines under this setting, with CatBoost achieving the best overall performance and XGBoost delivering the lowest MAE. These results confirm the robustness of gradient boosting approaches when forecasting industrial waste generation under fine-grained spatial aggregation.

The poor performance of ARIMA under H3 aggregation ($R^2 = 0.2633$) can be attributed to the violation of its core assumptions. ARIMA models assume stationarity and relatively continuous demand patterns, whereas industrial waste generation at fine spatial resolutions exhibits highly intermittent behavior with frequent zero-demand periods. This sparsity reduces the effectiveness of linear autoregressive models, highlighting the necessity of machine learning approaches capable of capturing non-linear and sparse demand patterns. The shift from traditional statistical baselines to machine learning models is particularly justified in the context of industrial waste generation forecasting. Waste generation processes often exhibit non-linear relationships and significant variability in both quantities and timing, meaning that the target variable cannot be adequately expressed as a simple linear function of explanatory variables. This behavior is particularly pronounced in industrial contexts characterized by intermittent and irregular generation patterns [5], which limits the effectiveness of linear statistical models such as ARIMA. Machine learning approaches are better suited to capture complex interactions between temporal, spatial, and operational variables.

The comparison between city-level and H3-based aggregation highlights an important trade-off between statistical performance and operational relevance. City-level aggregation produces smoother and more stable demand signals, resulting in higher explanatory power ($R^2 > 0.82$). However, such aggregation may mask localized spatial variability within urban

areas. Conversely, the H3-based hexagonal aggregation enables the identification of localized spatial patterns and potential operational hotspots. This finer spatial segmentation introduces greater sparsity and shorter historical time series per spatial unit, which naturally leads to lower R^2 values. Nevertheless, the reduction in absolute forecasting errors (MAE) at the H3 level remains operationally valuable, as it enables more precise route planning, resource allocation, and localized decision-making in reverse logistics systems.

4.5.3 Comparative Analysis and Operational Insights

The results reveal a fundamental trade-off between predictive accuracy and spatial resolution. City-level aggregation delivers highly stable demand patterns and superior forecasting performance, making it ideal for capacity planning and high-level decision-making. In contrast, H3-based aggregation enables localized spatial analysis and hotspot identification, valuable for route planning and service differentiation despite reduced predictability. These results demonstrate that spatial aggregation should be selected based on operational objectives rather than solely predictive performance, suggesting that combining coarse and fine spatial forecasting approaches can provide complementary decision support in industrial reverse logistics systems.

4.6 Conclusion

This paper presents an ML-based framework for industrial waste generation forecasting in reverse logistics, specifically examining the impact of spatio-temporal aggregation on predictive performance. Using real-world data from used motor oil container collections, classical statistical models were compared against advanced ML approaches across different spatial representations. A biweekly temporal aggregation effectively reduced demand intermittency while maintaining operational relevance. Experimental results confirmed that gradient boosting models substantially outperformed traditional methods. City-level aggregation delivered the highest predictive accuracy, whereas H3-based hexagonal aggregation enabled localized spatial analysis despite increased variability and lower predictability. These findings reveal a fundamental trade-off between forecasting accuracy and spatial resolution, demonstrating that aggregation choices must align with operational objectives rather than being treated as purely technical decisions. Future work will incorporate meteorological variables. For example, temperature variations can influence motor oil viscosity and industrial activity levels, potentially affecting waste generation volumes and collection efficiency. Integrating such environmental variables may affect forecasting accuracy and the operational robustness of reverse logistics planning.

CHAPITRE 5 ARTICLE 2 : DYNAMIC SPATIAL SMOOTHING BASED ON PRODUCTION PROFILE SIMILARITY FOR INDUSTRIAL WASTE FORECASTING

Submitted on March 10, 2026 and Published in ICTLE 2026 14th International Conference on Traffic and Logistics Engineering

Kso, A., Keivanpour, S., Frayret, J (2026, March 10)

Abstract : Industrial waste generation forecasting is a key enabler of reverse logistics operations and needs efficient resource planning. Although recent studies have shown that machine learning models can be effective for spatio-temporal waste forecasting, their performance often degrades in spatial units with sparse data or highly volatile historical patterns. This paper introduces a spatio-temporal smoothing framework designed to enhance forecast accuracy by exploiting both behavioral similarity and spatial proximity between geographic regions. The approach operates on waste collection data aggregated using the H3 hexagonal grid system at resolution 6 and bi-weekly time intervals. Initial forecasts are produced with a CatBoost regression model and subsequently refined through similarity-weighted aggregation of neighboring hexagons. The smoothing mechanism combines temporal production profiles with geographic adjacency to stabilize predictions in volatile areas. Weather variables such as temperature, precipitation, and snowfall are incorporated via a nearest-station mapping strategy to assess their impact on forecasting performance. Experimental results on a real-world dataset of industrial oil container collections across Canada demonstrate that the proposed smoothing framework enhances the overall robustness of the forecast. The method reduces extreme prediction errors, as reflected by a lower RMSE, and slightly improves the coefficient of determination (R^2), while maintaining comparable mean absolute error levels. Overall, the proposed framework offers a scalable and effective solution for enhancing waste generation forecasts and supports more informed decision-making in reverse logistics systems.

Keywords : Reverse logistics, waste forecasting, weather, spatial smoothing, H3 grid, machine learning, CatBoost, spatio-temporal modeling, industrial waste.

5.1 Introduction

Industrial waste collection and recycling play a critical role in supporting circular economy systems and reducing environmental impact. Accurate forecasting of waste generation is

essential for optimizing reverse logistics operations, including transportation planning, fleet management, and resource allocation. However, industrial waste generation exhibits strong spatial and temporal variability, particularly in applications such as motor oil container recovery, making accurate prediction challenging.

Recent advances in machine learning have significantly improved forecasting performance for complex spatio-temporal systems. Gradient boosting methods, particularly CatBoost, have demonstrated strong predictive capabilities due to their ability to capture nonlinear relationships and effectively handle categorical and heterogeneous features [44]. CatBoost introduces ordered boosting and specialized encoding techniques to reduce prediction bias and improve generalization performance, making it particularly suitable for industrial forecasting applications involving spatial and temporal identifiers [45]. In parallel, spatial aggregation has emerged as an effective strategy for structuring geospatial data and improving forecasting stability. The H3 hierarchical hexagonal spatial indexing system provides a scalable and consistent representation of geographic space by partitioning regions into hexagonal cells with well-defined neighborhood relationships [46, 47]. Hexagonal grids offer improved spatial continuity and more uniform neighbor relationships compared to square grids, enabling more effective modeling of spatial dependencies. In our previous work, industrial waste generation was aggregated using the H3 grid at resolution 6, and CatBoost regression was applied to generate baseline forecasts. While this approach achieved strong predictive performance, substantial variability remained across spatial units, particularly in regions characterized by sparse data or high volatility.

These limitations highlight the importance of incorporating spatial dependencies and behavioral similarity between geographic regions. Spatial autocorrelation is a well-established phenomenon in geospatial systems, where neighboring regions often exhibit similar temporal dynamics due to shared environmental conditions, operational characteristics, and geographic proximity. Ignoring these relationships may lead to unstable or suboptimal predictions, particularly in fine-grained spatial forecasting settings.

In addition to spatial relationships, environmental conditions may influence industrial waste generation patterns. Previous studies have demonstrated the impact of meteorological variables, including temperature and precipitation, on forecasting applications such as energy demand prediction and environmental modeling [48], [49]. Integrating weather data into machine learning forecasting models may therefore provide additional explanatory information and improve predictive performance, particularly for outdoor industrial activities. While machine learning models can effectively capture temporal patterns within each spatial unit, they typically treat observations as independent samples during training. As a result, spa-

tial dependencies and behavioral similarities between neighboring regions are not explicitly modeled. In fragmented and heterogeneous industrial waste systems, such spatial interactions may contain valuable information about shared operational patterns. Ignoring these relationships may therefore lead to unstable predictions in spatial units with sparse or highly volatile historical data. This motivates the development of a spatially adaptive smoothing mechanism capable of leveraging behavioral similarity across hexagonal units. To address these challenges, this study proposes a similarity-based spatio-temporal smoothing framework to enhance industrial waste forecasting accuracy. First, baseline predictions are generated using a CatBoost regression model applied to H3 spatial units at resolution 6. Meteorological variables, including temperature, snowfall, and precipitation, are integrated using Environment and Climate Change Canada station data mapped to spatial hexagons. Finally, a similarity-based spatial smoothing method is applied as a post-processing step, where predictions are refined using similarity-weighted information from neighboring hexagons derived from historical behavioral profiles. This study makes several key contributions by proposing a similarity-based spatial smoothing framework for industrial waste forecasting, integrating meteorological variables mapped to spatial hexagons via a nearest-station approach, and incorporating behavioral similarity profiles to enhance prediction stability. It is evaluated on a large-scale real-world industrial waste dataset and demonstrates improved forecasting accuracy compared to baseline machine learning models.

This paper is structured as follows : Section II reviews related work, Section III presents the proposed methodology, Section IV details and discusses the experimental results and Section V concludes the paper.

5.2 Literature Review

Spatial forecasting methods have been widely studied in logistics, transportation, and environmental applications. Classical spatial econometric models, such as spatial lag models, incorporate spatial dependencies using geographic proximity [50]. More recently, hierarchical spatial indexing systems such as the H3 hexagonal grid have been used to represent geographic areas consistently and efficiently for large-scale predictive modeling [51].

Machine learning models have significantly improved forecasting performance in spatio-temporal contexts. Gradient boosting and deep learning approaches have demonstrated strong predictive accuracy across multiple domains, including demand forecasting and waste generation prediction [52], [53]. However, most existing approaches do not explicitly exploit behavioral similarity between spatial units.

To address this limitation, several studies have proposed spatial smoothing techniques that incorporate information from neighboring regions. These methods improve forecast stability but often rely solely on geographic proximity and do not consider similarity in temporal generation patterns [54].

Table 5.1 summarizes the main spatial forecasting and smoothing approaches and highlights their limitations.

TABLE 5.1 Summary of spatial forecasting and similarity-based smoothing approaches

Reference	Application	Method and limitation
Anselin (1988)	Spatial statistics	<i>Method</i> : Spatial weighted averaging in spatial lag models. <i>Limitation</i> : Uses geographic distance only.
Brodsky <i>et al.</i> (2018)	Spatial indexing (H3)	<i>Method</i> : Hierarchical hexagonal spatial indexing. <i>Limitation</i> : Primarily used for spatial aggregation.
Wang <i>et al.</i> (2020)	Spatio-temporal forecasting	<i>Method</i> : Correlation-based similarity modeling. <i>Limitation</i> : Limited integration with machine learning prediction models.
Li <i>et al.</i> (2021)	Taxi demand forecasting	<i>Method</i> : Neighbor smoothing using weighted k-NN. <i>Limitation</i> : Does not consider behavioral similarity of production patterns.
Fildes <i>et al.</i> (2019)	Demand forecasting	<i>Method</i> : Machine learning and statistical forecasting models. <i>Limitation</i> : Does not explicitly model spatial interactions.
Bandara <i>et al.</i> (2020)	Time-series forecasting	<i>Method</i> : Deep learning forecasting models. <i>Limitation</i> : Does not capture spatial dependencies.
This study	Industrial waste forecasting	<i>Method</i> : CatBoost combined with similarity-based spatial smoothing. <i>Contribution</i> : Integrates spatial proximity and behavioral similarity to improve prediction accuracy.

As shown in Table 5.1, existing approaches either rely on geographic proximity alone or do not integrate behavioral similarity into machine learning forecasts. While prior studies have incorporated spatial proximity through weighted averaging or neighborhood-based smoothing, most approaches rely only on geographic distance without explicitly modeling behavioral

similarity in temporal production dynamics. Furthermore, existing machine learning forecasting frameworks often treat spatial units independently, limiting robustness in heterogeneous and volatile environments. The proposed framework bridges this gap by integrating dynamic behavioral similarity with spatial neighborhood structure in a post-processing forecasting setting. Recent advances in spatio-temporal forecasting have also explored deep learning approaches such as Graph Neural Networks (GNNs) to model spatial dependencies. While these models can effectively capture complex graph structures, they often require dense spatial connectivity and large training datasets. In industrial waste forecasting contexts characterized by sparse and heterogeneous spatial observations, tree-based ensemble models such as CatBoost provide a more computationally scalable and interpretable alternative [55], [56].

5.3 Methodology

5.3.1 Proposed Similarity-Based Forecasting Framework

Figure 5.1 illustrates the overall forecasting framework proposed in this study. The approach follows a two-stage strategy designed to improve the robustness of industrial waste generation forecasts.

In the first stage, a baseline machine learning model is used to generate initial predictions for each spatial unit and time period. Waste collection data are spatially aggregated using the H3 hierarchical hexagonal grid system and temporally aggregated into bi-weekly intervals. A CatBoost regression model is then trained to predict the waste generation level for each hexagon at each time step.

In the second stage, the baseline predictions are refined through a similarity-based spatial smoothing mechanism applied as a post-processing step. Instead of treating spatial units independently, this step incorporates information from neighboring hexagons that exhibit similar production dynamics. By leveraging behavioral similarity between spatial units, the proposed framework stabilizes predictions in regions characterized by sparse observations or highly volatile generation patterns.

This two-stage approach allows spatial interactions to be incorporated without modifying or retraining the underlying machine learning model, providing a scalable and computationally efficient forecasting strategy.

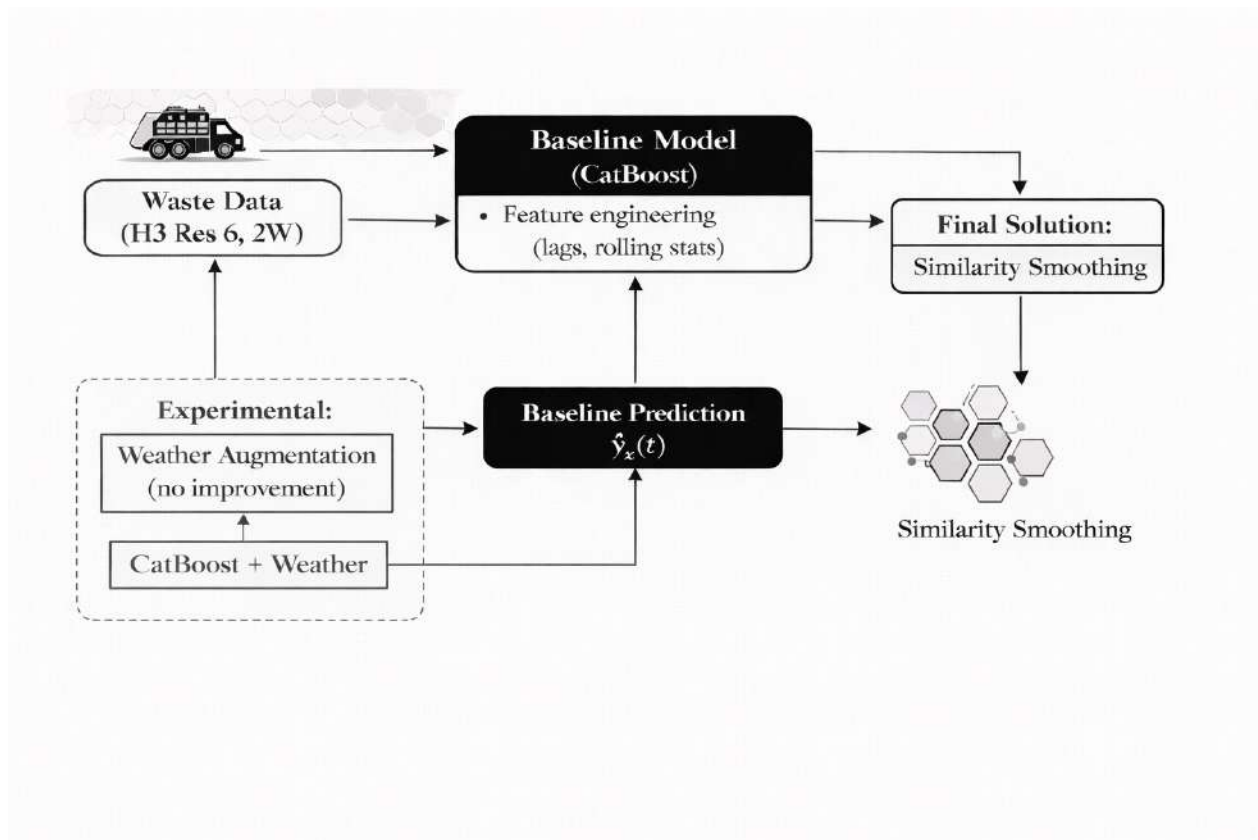


FIGURE 5.1 Overview of the proposed dynamic similarity-based smoothing framework.

5.3.2 Weather Feature Integration

Environmental conditions may influence industrial waste generation patterns, particularly in operational contexts where activities are affected by weather conditions. Weather variables were therefore included as potential explanatory factors in the forecasting framework. To evaluate the potential impact of environmental variables on forecasting performance, meteorological data from Environment and Climate Change Canada (ECCC) were integrated into the forecasting pipeline [57].

Each H3 hexagon was associated with the nearest available weather station using centroid-based geographic distance. When the closest station lacked sufficient observations, a fallback mechanism selected the nearest station with valid measurements.

Daily weather observations were aggregated using a rolling 14-day window in order to align with the bi-weekly temporal aggregation used for the waste generation dataset. For each hexagon x and time period t , the following features were computed :

- Mean air temperature over the preceding 14 days,
- Cumulative precipitation over the preceding 14 days,
- Cumulative snowfall over the preceding 14 days.

To prevent temporal information leakage, all rolling statistics were computed using only past observations through a one-period temporal shift. The resulting weather features were then merged with the waste generation dataset using matching spatial identifiers and time periods, producing a unified spatio-temporal dataset for model training and evaluation.

5.3.3 Dynamic Behavioral Profile Construction

Industrial waste generation exhibits heterogeneous and evolving temporal dynamics across spatial units. To capture these patterns, behavioral profiles were constructed for each hexagon using rolling statistical indicators computed over a historical window of L periods.

For each hexagon x at time t , behavioral features were computed using past model predictions $\hat{y}_x(i)$ for $i < t$ rather than raw observations $y_x(i)$.

Using model predictions provides a smoothed representation of the underlying generation patterns, which is particularly important in this dataset due to the high intermittency and sparsity of observed waste generation in several spatial units. In many hexagons, raw observations contain frequent zero values and irregular collection intervals, which can introduce noise in similarity measurements.

Using the baseline CatBoost predictions, the constructed behavioral profiles capture the

expected production dynamics learned by the forecasting model rather than the short-term observed noise. This approach improves the stability of similarity estimates between spatial units while maintaining temporal consistency.

To avoid temporal leakage, only past predictions $\hat{y}_x(i)$ with $i < t$ were used when computing the rolling statistics.

$$\mu_x(t) = \frac{1}{L} \sum_{i=t-L}^{t-1} \hat{y}_x(i) \quad (5.1)$$

where $\hat{y}_x(i)$ represents the predicted waste generation at period i .

The corresponding rolling standard deviation is computed as :

$$\sigma_x(t) = \sqrt{\frac{1}{L} \sum_{i=t-L}^{t-1} (\hat{y}_x(i) - \mu_x(t))^2} \quad (5.2)$$

These rolling statistics summarize the local temporal dynamics of waste generation and capture the variability and volatility of production patterns within each spatial unit. Rolling-window statistics are widely used in time-series modeling to represent evolving temporal behavior while preserving temporal causality [18].

In addition to the rolling mean and standard deviation, the rolling median was also computed to improve robustness against extreme variations. All profile features were calculated using temporally shifted observations in order to prevent information leakage and preserve the causal structure of the forecasting framework.

5.3.4 Similarity Computation and Spatial Smoothing

Spatial neighbors were identified using the H3 k -ring operator. For each hexagon x , the neighborhood set $N(x)$ contains all hexagons located within k hexagonal steps (with $k = 1$ or $k = 2$).

Behavioral similarity between two hexagons x and v at time t is computed using the Euclidean distance between their standardized profile vectors :

$$s_{xv}(t) = \exp \left(-\frac{\|p_x(t) - p_v(t)\|}{\tau} \right) \quad (5.3)$$

where $p_x(t)$ denotes the behavioral profile of hexagon x , $\|\cdot\|$ the Euclidean norm, and $\tau > 0$ is a temperature parameter controlling sensitivity to profile differences. This exponential

similarity function corresponds to a radial basis function (RBF) kernel commonly used in pattern recognition and similarity modeling [58].

To reduce noise and prevent the inclusion of weakly related spatial units, only the Top- K most similar neighbors are retained, forming the set $N_K(x)$.

Similarity weights are normalized like that :

$$\sum_{v \in N_K(x)} w_{xv}(t) = 1 \quad (5.4)$$

The parameter k controls the spatial neighborhood size. As illustrated in Fig 5.2, When $k = 1$, the neighborhood includes the six immediate adjacent hexagons. When $k = 2$, the neighborhood expands to include both the first and second rings, resulting in a total of 18 neighboring hexagons (6 in the first ring and 12 in the second ring). This hierarchical structure allows the model to incorporate spatial information at multiple scales [58].

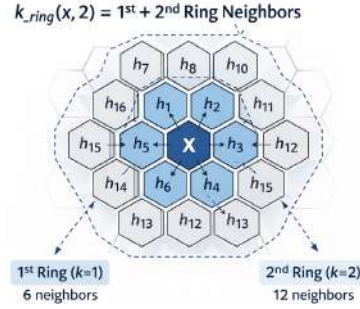


FIGURE 5.2 Illustration of the H3 k -ring neighborhood structure.

5.3.5 Similarity-Based Spatial Smoothing

The final prediction is computed as a weighted combination of the original CatBoost prediction and the similarity-weighted average of neighboring predictions :

$$\hat{y}_x^{\text{smooth}}(t) = (1 - \alpha)\hat{y}_x(t) + \alpha \sum_{v \in N_K(x)} w_{xv}(t)\hat{y}_v(t) \quad (5.5)$$

This formulation is conceptually related to spatial lag modeling and weighted spatial averaging approaches in spatial econometrics [50]. where $\hat{y}_x(t)$ is the baseline prediction, $N_K(x)$ is the set of Top- K most similar neighbors, $w_{xv}(t)$ are normalized similarity weights satisfying $\sum_{v \in N_K(x)} w_{xv}(t) = 1$, and $\alpha \in [0, 1]$ controls the smoothing strength.

The smoothing parameter α controls the trade-off between local predictive fidelity and spatial regularization.

When $\alpha \rightarrow 0$, the model reduces to the independent baseline.

When $\alpha \rightarrow 1$, predictions converge toward a spatially weighted average.

This proposed formulation acts as a spatial bias–variance balancing mechanism across hexagons, consistent with regularization principles in statistical learning theory [59]. Small values of α preserve local prediction characteristics, whereas larger values increase spatial averaging [60].

This post-processing approach enhances robustness by incorporating spatial behavioral similarity without retraining the underlying machine learning model.

5.3.6 Case Study : Industrial Oil Container Recovery in Canada

The proposed forecasting framework was evaluated using a real-world dataset of industrial oil container collection records collected across Canada between 2014 and 2024 developed in our previous work [44].

Collection events were spatially aggregated using the H3 hierarchical hexagonal grid system at resolution 6, producing approximately 3,868 unique H3 hexagonal spatial units. Temporally, observations were aggregated into bi-weekly intervals to align with operational planning cycles.

The resulting dataset exhibits strong spatial heterogeneity and temporal variability. Some hexagons present dense and stable waste generation patterns, while others exhibit intermittent or highly volatile behavior due to differences in industrial activity levels.

Figure ?? illustrates the spatial distribution of high-activity H3 hexagons during a representative bi-weekly period. The presence of localized clusters highlights the strong spatial heterogeneity of industrial waste generation patterns, with darker colors (from light blue to red) indicating higher waste generation levels.

After preprocessing and filtering incomplete observations, the final evaluation dataset contains approximately 60,730 spatio-temporal observations. This case study provides a realistic industrial context for evaluating the proposed similarity-based smoothing framework and assessing its ability to improve forecasting robustness in heterogeneous reverse logistics systems.

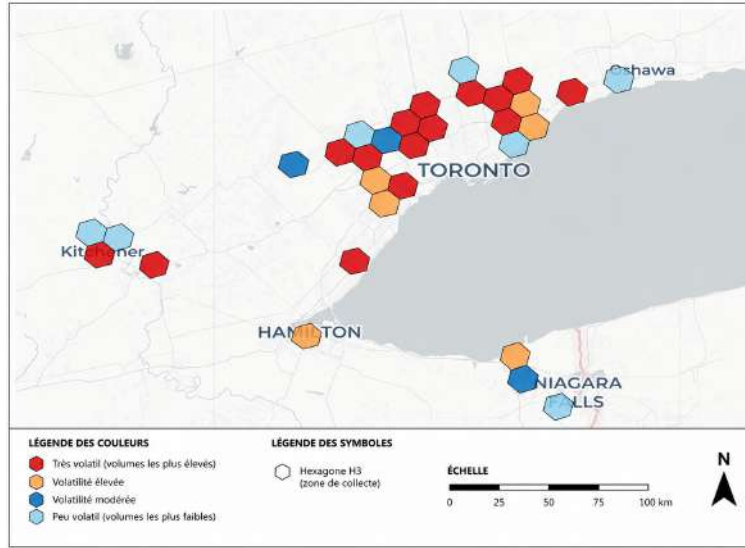


FIGURE 5.3 Spatial distribution of high-activity H3 hexagons for a representative bi-weekly period in Southern Ontario, Canada

5.4 Results and Discussion

5.4.1 Experimental Comparison of Forecasting Configurations

The analysis compares three applied configurations : (1) the baseline CatBoost model without spatial smoothing or weather information, (2) the CatBoost model with weather integration, and (3) the CatBoost model combined with the proposed similarity-based spatial smoothing.

Table 5.2 summarizes predictive performance across configurations.

TABLE 5.2 Predictive performance comparison across model configurations

Model	RMSE (kg)	MAE (kg)	R^2
CatBoost baseline	259.30	88.94	0.5867
CatBoost + Weather	262.90	90.80	0.5751
CatBoost + Spatial Smoothing	258.44	90.14	0.5894

The similarity-based smoothing configuration achieved the best overall performance, reducing RMSE and improving the coefficient of determination R^2 .

To further assess the impact of spatial smoothing on large prediction deviations, percentile absolute errors were computed for both models. The results are summarized in Table 5.3.

The results indicate that the proposed smoothing mechanism slightly reduces extreme predic-

TABLE 5.3 Percentile absolute error comparison

Metric	Baseline (kg)	Spatial Smoothing (kg)
90th percentile error	205.25	204.98
95th percentile error	310.88	310.51

tion errors. Although the improvements remain modest, they confirm that spatial smoothing mitigates large deviations while maintaining overall predictive stability.

5.4.2 Baseline Forecasting Performance

The baseline CatBoost model demonstrated strong predictive capability, confirming the suitability of gradient boosting methods for industrial waste forecasting. With an RMSE of 259.30 and $R^2 = 0.5867$, the model effectively captured spatial and temporal variability in waste generation across H3 hexagons.

These results are consistent with prior research demonstrating the effectiveness of boosting-based models in heterogeneous spatio-temporal datasets.

However, residual variability remained across spatial units, particularly in hexagons characterized by :

- Sparse historical observations,
- High volatility,
- Irregular production patterns.

This motivated the introduction of spatial behavioral smoothing.

5.4.3 Impact of Weather Variables

Weather variables : mean temperature, cumulative precipitation, and cumulative snowfall were incorporated to assess whether environmental conditions improve predictive accuracy.

Contrary to expectations, the inclusion of weather features degraded overall performance :

- RMSE increased from 259.30 to 262.90,
- MAE increased from 88.94 to 90.80,
- R^2 decreased from 0.5867 to 0.5751.

Several domain-specific factors may explain this result : Industrial waste generation appears to be primarily driven by operational schedules and maintenance cycles rather than short-term weather fluctuations. Moreover, bi-weekly temporal aggregation may smooth out meteo-

rological variability observable at finer time scales, while heterogeneity in industrial activity intensity likely dominates over environmental effects.

These findings suggest that in this reverse logistics context, operational dynamics dominates over weather influences.

Importantly, testing weather variables strengthens the methodological rigor of this study by empirically validating their limited contribution rather than assuming their relevance.

5.4.4 Impact of Similarity-Based Spatial Smoothing

The proposed similarity-based spatial smoothing mechanism achieved the best performance, with RMSE reduced to 258.44 (from 259.30), MAE slightly increasing to 90.14 (from 88.94), and R^2 improving to 0.5894 (from 0.5867).

Although the numerical improvement in RMSE approximately 0.33% appears moderate, the reduction is meaningful from an operational perspective. Because RMSE penalizes large deviations more heavily than MAE, the observed decrease indicates improved robustness against extreme forecasting errors. In reverse logistics systems, where significant underestimation can lead to inefficient routing, capacity shortages, or scheduling interruptions, mitigation of large errors is particularly valuable.

Unlike conventional spatial smoothing techniques based purely on geographic proximity, the proposed framework incorporates behavioral similarity derived from historical production dynamics. Neighboring hexagons are weighted according to similarity in temporal variation patterns rather than physical distance alone. This allows spatial information to be integrated in a structured and adaptive manner while preserving local predictive characteristics.

The smoothing procedure is implemented as a post-processing step applied to baseline machine learning predictions. This design avoids retraining the underlying model, maintains computational efficiency, and ensures scalability to large industrial datasets.

The slight increase in MAE suggests that the smoothing mechanism primarily attenuates large deviations rather than uniformly improving all prediction errors. This behavior reflects a spatial bias-variance trade-off, where regularization across neighboring units reduces extreme fluctuations while slightly smoothing localized variations.

Figures 5.4 and 5.5 illustrate two consecutive segments of the same highly volatile hexagonal unit time series for visualization clarity. The observed values are compared with the baseline CatBoost predictions and the proposed similarity-based smoothed predictions.

While the baseline model exhibits noticeable fluctuations, the similarity-based spatial smoo-

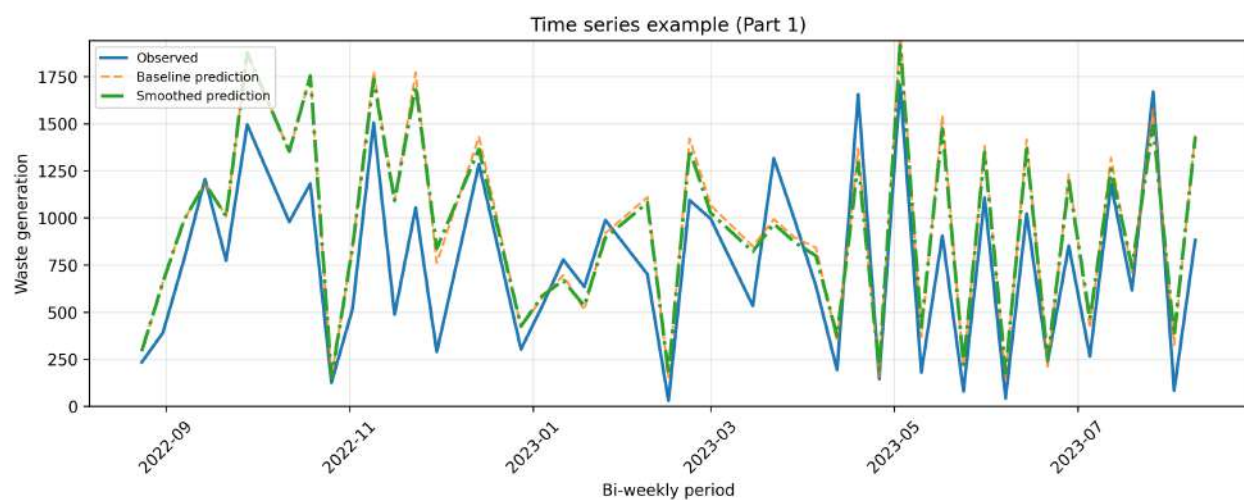


FIGURE 5.4 Time-series example (Part 1) for a highly volatile hexagonal unit

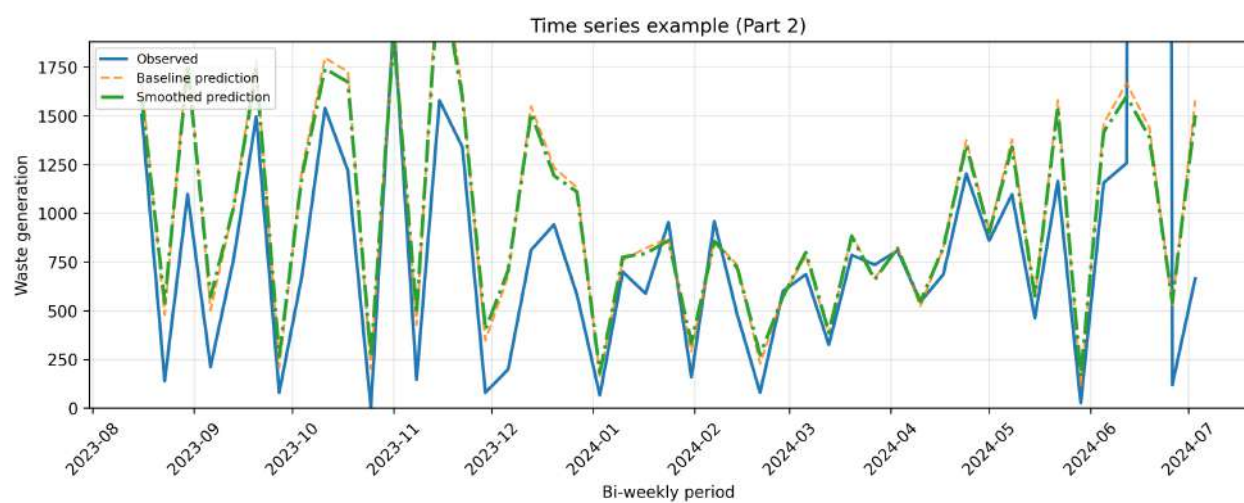


FIGURE 5.5 Time-series example (Part 2) for a highly volatile hexagonal unit

thing approach produces more stable and consistent forecasts, better capturing the underlying temporal dynamics.

Overall, the findings indicate that similarity-based spatial interaction provides complementary predictive information beyond independent hexagon-level machine learning forecasts, enhancing robustness in heterogeneous and volatile spatial environments.

5.5 Conclusion

This study proposed a similarity-based spatial smoothing framework to improve machine learning forecasts of industrial waste generation in reverse logistics systems. Building on an H3-based CatBoost forecasting model, the proposed method introduces a post-processing mechanism that incorporates behavioral similarity and spatial proximity between hexagonal units.

The experimental results demonstrated that while the integration of meteorological variables did not improve predictive performance, the similarity-based spatial smoothing approach achieved the best overall results. The method reduced RMSE and improved the coefficient of determination R^2 , indicating increased robustness and better handling of large prediction errors in volatile spatial units.

The findings suggest that industrial waste generation patterns in this context are primarily driven by operational and behavioral dynamics rather than short-term environmental conditions. Incorporating similarity-weighted neighbor information allows the model to exploit spatial behavioral dependencies without retraining the underlying machine learning algorithm. From a practical perspective, stabilizing prediction outputs can directly support operational decision-making in reverse logistics systems. More stable forecasts reduce the risk of large prediction errors that can lead to inefficient routing plans, unnecessary vehicle dispatching, or capacity mismatches during collection operations. In Industry 4.0 logistics environments, where data-driven planning systems increasingly support operational decisions, improving forecast robustness can contribute to more reliable route planning, better fleet utilization, and improved collection efficiency.

Future research may explore adaptive similarity mechanisms, graph-based learning approaches, and dynamic clustering strategies that directly integrate spatial relationships during model training. Extending the framework to other reverse logistics or industrial forecasting contexts also represents a promising direction.

CHAPITRE 6 CONCLUSION ET DISCUSSION GÉNÉRALE

6.1 Synthèse des travaux

La prévision de la génération des résidus industriels constitue un enjeu central pour l'amélioration de la performance des systèmes de logistique inverse. Dans un contexte où les organisations cherchent à optimiser les opérations de collecte et à soutenir les principes de l'économie circulaire, la capacité à anticiper les volumes de résidus générés devient un élément clé pour la prise de décision.

Les résultats obtenus dans ce mémoire montrent que les approches basées sur l'apprentissage automatique permettent d'améliorer significativement la précision des prévisions par rapport aux méthodes traditionnelles. En particulier, les modèles de type gradient boosting, notamment CatBoost, ont démontré les meilleures performances.

À l'échelle agrégée, les modèles développés atteignent un coefficient de détermination supérieur à 0.82, tout en réduisant significativement les erreurs de prévision (RMSE et MAE) par rapport aux approches classiques telles que les modèles ARIMA ou Croston.

À une échelle spatiale fine (H3 résolution 6), les modèles développés permettent de maintenir des performances prédictives robustes malgré une forte variabilité des données (coefficient de variation supérieur à 2). L'intégration d'un mécanisme de lissage spatial basé sur la similarité comportementale permet d'améliorer davantage les performances, en réduisant le RMSE et en augmentant le coefficient de détermination.

En revanche, l'intégration des variables météorologiques n'a pas permis d'améliorer significativement les performances prédictives, suggérant que la génération des résidus industriels est davantage influencée par des facteurs opérationnels et comportementaux que par des conditions environnementales à court terme.

Dans l'ensemble, ces résultats confirment la pertinence d'une approche combinant apprentissage automatique, structuration spatiale et analyse comportementale pour améliorer la prévision des résidus industriels. Ces travaux démontrent ainsi qu'une meilleure modélisation des dynamiques spatio-temporelles constitue un levier clé pour renforcer la performance et la durabilité des systèmes de logistique inverse.

6.2 Implications de la recherche

Les résultats obtenus présentent plusieurs implications importantes, tant sur le plan scientifique que sur le plan industriel et opérationnel.

Implications académiques et scientifiques

Sur le plan académique et scientifique, cette recherche contribue à la littérature sur la prévision dans les systèmes de logistique inverse en proposant un cadre de modélisation intégrant agrégation spatio-temporelle, apprentissage automatique et lissage spatial basé sur la similarité comportementale.

Elle met en évidence l'importance du choix de la granularité spatiale dans la performance des modèles prédictifs, en lien avec la problématique du Modifiable Areal Unit Problem (MAUP). Ce travail montre également l'intérêt d'intégrer des approches comportementales pour améliorer les modèles de prévision, au-delà des méthodes reposant uniquement sur la proximité géographique.

Enfin, cette recherche ouvre des perspectives pour le développement de modèles intégrant explicitement les dépendances spatiales dans le processus d'apprentissage, ainsi que pour l'exploration de nouvelles variables explicatives adaptées aux systèmes de logistique inverse.

Implications industrielles et opérationnelles

Sur le plan industriel et opérationnel, les approches proposées offrent des outils d'aide à la décision permettant d'améliorer la planification des opérations de collecte dans les systèmes de récupération des résidus industriels.

Des prévisions plus fiables permettent notamment d'anticiper avec plus de précision les volumes à collecter, ce qui facilite la planification des tournées, l'allocation des ressources et la gestion des capacités logistiques. Une meilleure anticipation contribue directement à réduire les déplacements inutiles, à limiter les situations de surcharge ou de sous-utilisation des véhicules, et à améliorer la qualité du service offert aux clients.

Par ailleurs, la réduction des déplacements inutiles entraîne une diminution des émissions de gaz à effet de serre associées aux activités de transport et limite la saturation des centres de tri. Ces bénéfices renforcent le lien entre amélioration des performances opérationnelles et objectifs de durabilité dans une perspective d'économie circulaire.

Ainsi, l'intégration de modèles prédictifs robustes dans les systèmes logistiques constitue un levier concret pour améliorer la performance tout en réduisant l'impact environnemental.

Au-delà des gains opérationnels immédiats, ces résultats s’inscrivent dans une dynamique plus large liée à l’Industrie 4.0 et à la transition vers des systèmes manufacturiers circulaires. L’intégration de données hétérogènes (opérationnelles, spatiales et contextuelles) dans des modèles prédictifs avancés permet de renforcer la prise de décision basée sur les données (*data-driven decision making*). Cette approche contribue à la mise en place de systèmes logistiques plus intelligents, adaptatifs et durables, capables de mieux répondre aux enjeux de performance et de circularité des ressources.

6.3 Limites

Malgré les contributions apportées, certaines limites doivent être reconnues.

Tout d’abord, cette étude repose sur un cas industriel spécifique lié à la collecte de bidons d’huile moteur au Canada. Bien que les méthodes proposées soient génériques, leur application à d’autres types de résidus industriels ou à d’autres contextes géographiques pourrait nécessiter des ajustements.

Par exemple, certains secteurs industriels présentent des dynamiques de génération très différentes. Dans l’industrie aéronautique, les cycles de maintenance sont plus longs et moins réguliers, ce qui peut entraîner des flux de résidus plus intermittents. De même, dans le cas des déchets électroniques (*e-waste*), la génération dépend fortement des cycles de remplacement technologique et des comportements des consommateurs, ce qui introduit des variations difficiles à capturer avec les mêmes variables explicatives.

Ensuite, les modèles développés reposent principalement sur des variables issues des données opérationnelles historiques. L’intégration d’autres variables contextuelles, telles que les conditions économiques, les politiques environnementales ou certaines caractéristiques sectorielles, pourrait permettre d’enrichir les modèles et d’améliorer leur capacité explicative.

Bien que différentes granularités temporelles (hebdomadaire, bi-hebdomadaire et mensuelle) aient été explorées dans les phases initiales de cette recherche, l’analyse a été poursuivie en se concentrant sur une agrégation bi-hebdomadaire, jugée comme offrant le meilleur compromis entre stabilité et performance prédictive.

Cependant, ce choix introduit une limitation importante d’un point de vue opérationnel. En effet, l’agrégation bi-hebdomadaire tend à lisser les variations à court terme et réduit la capacité du modèle à capturer des fluctuations rapides de la génération des résidus. Cette contrainte peut limiter l’utilisation des modèles pour des décisions en temps quasi réel, telles que l’ajustement dynamique des tournées ou la gestion quotidienne des opérations de collecte.

Toutefois, ce choix implique une limitation, dans la mesure où une approche hiérarchique

multi-échelles n'a pas été développée. De telles approches, combinant plusieurs niveaux d'agrégation temporelle, pourraient permettre de mieux capturer les dynamiques à différentes échelles et d'améliorer la robustesse des prévisions. L'exploration de modèles hiérarchiques ou multi-résolution constitue ainsi une perspective intéressante pour de futurs travaux.

Par ailleurs, l'approche de lissage spatial proposée est appliquée comme une étape de post-traitement des prédictions du modèle de base. Bien que cette stratégie présente l'avantage d'être simple et flexible, des approches intégrant directement les dépendances spatiales dans le processus d'apprentissage pourraient permettre d'obtenir des gains de performance supplémentaires.

Enfin, bien que certaines variables météorologiques aient été explorées, leur impact s'est avéré limité dans ce contexte spécifique. Cela suggère que d'autres facteurs explicatifs, notamment liés aux opérations industrielles, jouent un rôle plus déterminant dans la génération des résidus.

En outre, l'intégration de variables économiques pourrait constituer une piste intéressante pour améliorer les performances prédictives dans certains contextes industriels. Des variables telles que les prix des carburants, les fluctuations des matières premières, les indicateurs économiques sectoriels ou les niveaux d'activité industrielle pourraient influencer indirectement la génération des résidus et les besoins de collecte. L'exploration de ces facteurs exogènes représente ainsi une avenue prometteuse pour enrichir les modèles développés dans ce mémoire.

6.4 Recherches futures

Les résultats obtenus dans ce mémoire ouvrent plusieurs perspectives de recherche visant à approfondir et à étendre les approches proposées.

Une première direction consiste à enrichir les modèles prédictifs par l'intégration de variables contextuelles supplémentaires, telles que des indicateurs économiques, des caractéristiques industrielles locales ou des données opérationnelles plus fines. Bien que les variables météorologiques aient été explorées dans ce travail, leur contribution s'est révélée limitée, suggérant que d'autres facteurs exogènes pourraient jouer un rôle plus déterminant dans la génération des résidus industriels.

Une seconde piste concerne le développement d'approches avancées de modélisation spatio-temporelle. En particulier, les modèles basés sur les graphes (Graph Neural Networks – GNNs) ou les réseaux neuronaux spatio-temporels pourraient permettre de capturer de manière plus explicite les interactions complexes entre unités géographiques, en exploitant la structure du

voisinage spatial définie par le système H3. Toutefois, ces approches introduisent un compromis important : bien qu’elles puissent améliorer la capacité de modélisation des dépendances spatiales, elles se font souvent au détriment de l’interprétabilité, qui constitue un avantage clé des modèles tabulaires tels que CatBoost.

Par ailleurs, l’exploration de méthodes de regroupement dynamique (clustering spatio-temporel) constitue une avenue prometteuse pour améliorer la robustesse des modèles prédictifs. De telles approches permettraient d’identifier automatiquement des groupes d’unités spatiales partageant des comportements similaires, facilitant ainsi l’adaptation des modèles aux différentes typologies de générateurs de résidus. L’intégration de variables économiques, telles que les prix des matières premières, les indicateurs industriels ou les fluctuations économiques sectorielles, constitue également une avenue prometteuse pour améliorer la capacité explicative des modèles prédictifs dans certains contextes industriels.

Enfin, une perspective particulièrement intéressante consiste à intégrer les modèles prédictifs développés dans des cadres d’aide à la décision en logistique. Les prévisions de génération de résidus pourraient ainsi être utilisées pour alimenter des modèles d’optimisation des tournées de collecte sous incertitude, notamment dans le cadre du problème de tournées de véhicules stochastique (Stochastic Vehicle Routing Problem – SVRP). Cette intégration permettrait de relier directement la prévision à la prise de décision opérationnelle, en améliorant simultanément l’efficacité logistique et la robustesse des plans de collecte.

En définitive, ce mémoire met en évidence le potentiel des approches intégrées combinant apprentissage automatique, structuration spatio-temporelle et mécanismes de régularisation spatiale pour répondre aux défis complexes de la logistique inverse. Ces travaux ouvrent la voie à des systèmes de prévision plus performants et mieux alignés avec les exigences des systèmes industriels modernes et durables.

RÉFÉRENCES

- [1] World Bank, “Solid waste management,” 2022, accessed : 2026-03-16. [En ligne]. Disponible : <https://www.worldbank.org/en/topic/urbandevelopment/brief/solid-waste-management>
- [2] Environment and Climate Change Canada, “Solid waste diversion and disposal,” 2024, accessed : 2026-03-16. [En ligne]. Disponible : <https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/solid-waste-diversion-disposal.html>
- [3] M. Lieder et A. Rashid, “Towards circular economy implementation : A comprehensive review in context of manufacturing industry,” *Journal of Cleaner Production*, vol. 115, p. 36–51, 2016.
- [4] F. Acerbi et M. Taisch, “A literature review on circular economy adoption in the manufacturing sector,” *Journal of Cleaner Production*, vol. 273, p. 123086, 2020.
- [5] S. Makridakis, E. Spiliotis et V. Assimakopoulos, “Statistical and machine learning forecasting methods : Concerns and ways forward,” *PloS one*, vol. 13, n°. 3, p. e0194889, 2018.
- [6] D. S. Rogers et R. S. Tibben-Lembke, “An examination of reverse logistics practices,” *Journal of Business Logistics*, vol. 22, n°. 2, p. 129–148, 2001.
- [7] S. Agrawal, R. K. Singh et Q. Murtaza, “A literature review and perspectives in reverse logistics,” *Resources, Conservation and Recycling*, vol. 97, p. 76–92, 2015.
- [8] K. Bayramov, “The role of software in reverse logistics and effect on oily waste management,” *Journal of Transport and Supply Chain Management*, vol. 17, p. a941, 2023.
- [9] O. Ouro-Salim, “Urban mining of e-waste management globally : Literature review,” *Cleaner Waste Systems*, vol. 9, p. 100162, 2024.
- [10] O. Bhowmik, S. Chowdhury, J. H. Ashik, G. I. Mahmud, M. M. Khan et N. U. I. Hossain, “Application of artificial intelligence in reverse logistics : A bibliometric and network analysis,” *Supply Chain Analytics*, vol. 7, p. 100076, 2024.
- [11] G. Ghiani, G. Laporte et R. Musmanno, *Introduction to logistics systems management*. John Wiley & Sons, 2014.
- [12] F. Fernandes de Souza *et al.*, “A forecasting model based on arima and artificial neural networks for end-of-life vehicles,” *Journal of Cleaner Production*, vol. 337, p. 130441, 2022.

- [13] J. Shao et al., “Net benefits change of waste photovoltaic recycling in china : Projection of waste based on multiple factors,” *Journal of Cleaner Production*, vol. 389, p. 136051, 2023.
- [14] R. J. Hyndman et J. E. Boylan, “Stochastic models underlying croston’s method for intermittent demand forecasting,” *International Journal of Forecasting*, vol. 24, n°. 3, p. 389–402, 2008.
- [15] F. Ducharme, “Improving demand forecasting for customers with missing information,” Mémoire de maîtrise, École Polytechnique de Montréal, Montréal, QC, 2024.
- [16] R. Fildes, S. Ma et S. Kolassa, “Retail forecasting : Research and practice,” *International Journal of Forecasting*, vol. 35, n°. 1, p. 1–14, 2019.
- [17] K. Bandara, C. Bergmeir et S. Smyl, “Forecasting across time series databases using recurrent neural networks,” *International Journal of Forecasting*, vol. 36, n°. 4, p. 1388–1406, 2020.
- [18] R. J. Hyndman et G. Athanasopoulos, *Forecasting : Principles and Practice*, 2^e éd. Melbourne, Australia : OTexts, 2018.
- [19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye et T.-Y. Liu, “Lightgbm : A highly efficient gradient boosting decision tree,” dans *Advances in Neural Information Processing Systems*, 2017.
- [20] S. Chen *et al.*, “Spatio-temporal signal recovery under diffusion-induced smoothness and temporal correlation priors,” *IET Signal Processing*, 2021.
- [21] L. Anselin, *Spatial Econometrics : Methods and Models*. Dordrecht, The Netherlands : Kluwer Academic Publishers, 1988.
- [22] X. Wang, Y. Liu et H. Wang, “Spatio-temporal forecasting using correlation-based similarity measures,” *IEEE Access*, vol. 8, p. 148 389–148 401, 2020.
- [23] Y. Li, Y. Zheng, H. Zhang et L. Chen, “Traffic prediction in a bike-sharing system,” dans *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015.
- [24] Z. K. Moon *et al.*, “Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem,” *International Journal of Health Geographics*, 2019.
- [25] A. S. Fotheringham, C. Brunsdon et M. Charlton, *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*. Wiley, 2002.
- [26] N. Soares, P. Walther et M. Werner, “Experiments on geospatial data modelling for long-term trajectory prediction of aircrafts,” *AGILE : GIScience Series*, vol. 6, p. 46, 2025.

- [27] J. Zhang, Y. Zheng et D. Qi, “Deep spatio-temporal residual networks for citywide crowd flows prediction,” dans *AAAI Conference on Artificial Intelligence*, 2017.
- [28] S. Aghabozorgi, A. S. Shirkhorshidi et T. Y. Wah, “Time-series clustering : A decade review,” *Information Systems*, vol. 53, p. 16–38, 2015.
- [29] L. Ding, T. Wang et P. W. Chan, “Forward and reverse logistics for circular economy in construction : A systematic literature review,” *Journal of Cleaner Production*, vol. 388, p. 135981, 2023.
- [30] K. Bayramov, “The role of software in reverse logistics and effect on oily waste management,” *Journal of Transport and Supply Chain Management*, vol. 17, p. a941, 2023.
- [31] M. Ouro-Salim, “Urban mining of e-waste management globally : A literature review,” *Cleaner Waste Systems*, vol. 7, p. 100116, 2024.
- [32] J. Shao *et al.*, “Net benefits change of waste photovoltaic recycling in china : Projection of waste based on multiple factors,” *Journal of Cleaner Production*, vol. 389, p. 136051, 2023.
- [33] O. Daramola *et al.*, “Multivariate time series forecasting of municipal solid waste,” *Waste Management*, 2024.
- [34] O. Bhowmik *et al.*, “Application of artificial intelligence in reverse logistics : A bibliometric and network analysis,” *Algorithms*, vol. 18, n^o. 6, p. 650, 2025.
- [35] S. Kimiagari, S. Keivanpour et M. Haverila, “Developing a high-performance clustering framework for global market segmentation and strategic profiling,” *Journal of Business Research*, vol. 126, p. 94–110, 2021.
- [36] Y. Xu, X. Chen, Z. Li et Y. Zhang, “A h3-based spatial framework for large-scale urban mobility analysis,” *Transportation Research Part C : Emerging Technologies*, vol. 139, p. 103655, 2022.
- [37] K. Fang, Z. He, X. Wang et Y. Lin, “Hexagonal grid-based spatial clustering for urban service analysis,” *IEEE Access*, vol. 9, p. 118 247–118 259, 2021.
- [38] A. Boujamza, S. L. Elhaq et A. Loukili, “Ensemble forecasting for predicting petroleum products traffic in oil terminal,” dans *2023 14th International Conference on Intelligent Systems : Theories and Applications (SITA)*. IEEE, 2023, p. 1–6.
- [39] S. Natalia *et al.*, “A gis-based framework for scalable spatio-temporal prediction using hexagonal indexing,” *AGILE : GIScience Series*, vol. 6, p. 46, 2025.
- [40] A. Rostami-Tabar, A. Babai et R. Syntetos, “Demand forecasting for supply chains under temporal aggregation,” *European Journal of Operational Research*, vol. 281, n^o. 2, p. 291–304, 2019.

- [41] A. Kourentzes, I. Petropoulos et S. Trapero, “Improving forecasting by temporal aggregation and combination,” *International Journal of Forecasting*, vol. 38, n° 3, p. 1014–1031, 2022.
- [42] S. Sanguri, R. Babai et A. Syntetos, “Intermittent demand forecasting under temporal aggregation and obsolescence,” *International Journal of Production Research*, vol. 62, n° 4, p. 1152–1170, 2024.
- [43] L. Pasa, M. Rossi et F. Rizzoli, “Enhancing door-to-door waste collection forecasting through machine learning,” *Waste Management*, vol. 176, p. 105007, 2025.
- [44] A. Ksouma, S. Keivanpour, J.-M. Frayret et A. Arminia, “Optimizing reverse logistics : A spatio-temporal ml framework for industrial waste forecasting,” 2026, submitted for publication at ICTLE Conference.
- [45] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush et A. Gulin, “Catboost : Unbiased boosting with categorical features,” dans *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [46] B. Brodsky, D. Shemesh et R. K. Panicker, “Geospatial analytics using hierarchical hexagonal indexing,” *ISPRS International Journal of Geo-Information*, vol. 7, n° 9, p. 345, 2018.
- [47] D. Bousquin, M. Ward et A. Sampson, “Spatial clustering and prediction using hierarchical spatial indexing,” *Environmental Modelling & Software*, vol. 143, 2021.
- [48] R. Fildes, S. Ma et S. Kolassa, “Retail forecasting : Research and practice,” *International Journal of Forecasting*, vol. 35, n° 1, p. 1–14, 2019.
- [49] K. Bandara, C. Bergmeir et S. Smyl, “Forecasting across time series databases using recurrent neural networks,” *International Journal of Forecasting*, vol. 36, n° 4, p. 1388–1406, 2020.
- [50] L. Anselin, *Spatial Econometrics : Methods and Models*. Dordrecht, The Netherlands : Kluwer Academic Publishers, 1988.
- [51] X. Wang, Y. Liu et H. Wang, “Spatio-temporal forecasting using correlation-based similarity measures,” *IEEE Access*, vol. 8, p. 148 389–148 401, 2020.
- [52] T. Chen et C. Guestrin, “Xgboost : A scalable tree boosting system,” dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794.
- [53] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye et T. Liu, “Lightgbm : A highly efficient gradient boosting decision tree,” dans *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

- [54] Y. Li, Y. Zheng, H. Zhang et L. Chen, “Traffic prediction in a bike-sharing system,” dans *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015, p. 1–10.
- [55] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang et P. S. Yu, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, n^o. 1, p. 4–24, 2021.
- [56] B. Yu, H. Yin et Z. Zhu, “Spatio-temporal graph convolutional networks : A deep learning framework for traffic forecasting,” dans *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, 2018, p. 3634–3640.
- [57] Environment and Climate Change Canada, “Historical climate data,” 2024, government of Canada.
- [58] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York : Springer, 2006.
- [59] S. Geman, E. Bienenstock et R. Doursat, “Neural networks and the bias/variance dilemma,” *Neural Computation*, vol. 4, n^o. 1, p. 1–58, 1992.
- [60] T. Januschowski, J. Gasthaus, Y. Wang et D. Salinas, “Forecasting with spatio-temporal models,” *International Journal of Forecasting*, vol. 36, n^o. 3, p. 1039–1050, 2020.