

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

École doctorale : SCIENCES DU NUMÉRIQUE ET L'INGÉNIEUR

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

DOCTOR OF PHILOSOPHY OF DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

Discipline : INFORMATIQUE - COMPUTER SCIENCE

Présentée et soutenue publiquement par

Juliet Chebet MOSO

Le 4 Février 2022

**Approches d'exploration des flux de données dans les systèmes
de transport intelligents et l'agriculture de précision**
**Data Stream Mining Approaches in Intelligent Transportation
Systems and Precision Agriculture**

Jury :

M.	Francis ROUSSEAU	Professeur	URCA-CReSTIC	Président
M.	Henry NYONGESA	Professeur	DeKUT	Examineur
M.	Éric RENAULT	Professeur	LIGM	Rapporteur
M.	Ronald Waweru MWANGI	Professeur	JKUAT	Rapporteur
M.	Hacène FOUCHAL	Professeur	URCA-CReSTIC	Directeur de thèse
M.	Cyril DE RUNZ	Maître de conférences	BDTLN, LIFAT	Co-directeur de thèse
M.	Stéphane CORMIER	Maître de conférences	URCA-CReSTIC	Co-encadrant de thèse
M.	John Mwangi WANDETO	Maître assistant	DeKUT	Co-directeur de thèse

“Life is the art of drawing sufficient conclusions from insufficient premises.”

Samuel Butler

Abstract

In this thesis, we address the problem of analysing IoT data with a focus on anomaly detection in data streams and behaviour analysis. Unsupervised learning is highly preferred for real-life applications, especially in anomaly detection since there is a lot of data without labels in this scenario. We propose an Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP) technique. We define an unsupervised data-driven methodology and apply it in three case studies; detection of crop damage in crop dataset, application to GPS logs of combine harvesters and application to Cooperative Intelligent Transport System (C-ITS) messages. The focus is the identification of anomalies that can be linked to crop state/health during harvest, those that have an impact on harvest efficiency and those impacting road safety and efficiency. Based on our results, it is possible to link anomalies extracted to damaged crop state at the end of harvest. Also, we were able to detect deviant behaviour of combine-harvester and to identify anomalies on the roads. Therefore, anomaly detection could be integrated in the decision process of farm and road operators to improve harvesting efficiency, crop health, road safety and traffic flow.

Secondly, we considered the analysis of speed signatures generated from C-ITS messages with the aim of understanding driving behaviour evolution under a naturalistic driving environment. We have shown that with the application of segmentation and aggregate statistics, one is able to get a better understanding of general driving behaviour and infer information that relates to the road condition and traffic situation. Finally, we considered the trajectory-linking problem and applied it to C-ITS messages. Based on our analysis, it is possible to link trajectories to the generating users if other distinguishing attributes and background knowledge on generation of the messages are considered during similarity analysis.

Keywords: Anomaly detection, Data streams, Intelligent Transportation Systems, Smart farming, Traffic incident detection, Unsupervised learning.

Résumé court

Dans cette thèse, nous abordons le problème de l'analyse des données IoT en nous concentrant sur la détection des anomalies dans les flux de données et l'analyse des comportements. L'apprentissage non supervisé est intéressant pour les applications de la vraie vie, en particulier pour la détection des anomalies, car il y a beaucoup de données sans étiquettes dans ce scénario. Nous proposons une technique ELSCP (Enhanced Locally Selective Combination in Parallel outlier ensembles). Nous définissons une méthodologie non supervisée axée sur les données et nous l'appliquons à trois études de cas : la détection des dommages causés aux cultures dans un ensemble de données d'agriculture, l'application aux traces GPS des moissonneuses-batteuses et l'application aux messages issus des systèmes de transport intelligent coopératif (C-ITS). L'accent est mis sur l'identification des anomalies qui peuvent être liées à l'état ou à la santé des cultures pendant la récolte, celles qui ont un impact sur l'efficacité de la récolte et celles qui ont un impact sur la sécurité et le trafic routier. D'après nos résultats, il est possible de relier les anomalies extraites à l'état des cultures endommagées à la fin de la récolte. De même, nous avons été en mesure de détecter le comportement déviant de la moissonneuse-batteuse et d'identifier les anomalies sur les routes. Par conséquent, la détection des anomalies pourrait être intégrée dans le processus de décision des exploitants agricoles et routiers afin d'améliorer l'efficacité de la récolte, la santé des cultures, la sécurité routière et la fluidité du trafic.

Deuxièmement, nous avons considéré l'analyse des signatures de vitesse générées à partir des messages C-ITS dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturelle. Nous avons montré qu'avec l'application de la segmentation et des statistiques agrégées, on est capable d'obtenir une meilleure compréhension du comportement général de conduite et de déduire des informations relatives à l'état de la route et à la situation du trafic routier. Enfin, nous avons examiné le problème de la liaison de trajectoires et l'avons appliqué aux messages C-ITS. Suite à notre analyse, il est possible de relier les trajectoires aux utilisateurs qui les ont générées si d'autres attributs discriminants et des connaissances de base sur la génération des messages sont pris en compte pendant l'analyse de similarité.

Acknowledgements

The three-year journey to obtain this PhD degree would not have been feasible without the assistance of the following people:

I sincerely thank my thesis supervisors Stéphane Cormier, Hacène Fouchal, Cyril de Runz and John Mwangi Wandeto for all the immense support. In particular, for mentorship and guidance on background ideas which formed the theoretical foundation on which this work is built. The rigor and seriousness kept me on track at all times.

I would like to thank CReSTIC for welcoming me, providing resources and creating a favorable working environment during my stays in France.

I would like to thank my employer, Dedan Kimathi University of Technology (DeKUT), for creating a pleasant working atmosphere. I would like to express my heartfelt appreciation to the members of DeKUT's School of Computer Science and Information Technology for their invaluable support.

I am grateful to the French Embassy in Kenya for providing me with a scholarship opportunity. Their financial and social support enabled me to complete my thesis.

I would also like to thank the members of this thesis jury for having accepted to participate.

I would also like to thank Seçil, Brice and Ramzi with whom I had the pleasure to collaborate.

Most of all I thank the Lord for his blessings and my family for walking with me through this journey.

Table des matières

Résumé court	iv
Table des matières	vii
Introduction	ix
Contexte	ix
Contribution	xi
Plan de la thèse	xiii
Synthèse	xiv
État de l’art	xiv
Contexte et méthodes	xvii
Algorithme LSCP amélioré pour la détection des anomalies dans les données IoT	xviii
Signatures de données à partir de données IoT	xix
Liaison des données IoT	xx
Conclusion et perspectives	xxi
Conclusion	xxi
Limites	xxiii
Perspectives	xxiv

Résumé étendu

Introduction

Contexte

L'Internet des objets (IoT) automatise l'exécution d'actions intelligentes basées sur des données provenant d'appareils connectés en s'appuyant sur des technologies essentielles telles que la collecte automatisée de données par capteurs et l'analyse de données massives dans le nuage. Cette automatisation permet un large éventail d'applications pratiques dans le monde réel, notamment le transport intelligent, l'agriculture intelligente, les villes intelligentes, etc. L'incorporation de l'IoT dans la vie quotidienne, a entraîné la collecte de quantités massives de données. Les flux de données sont des séquences vastes, continues et non limitées de données qui arrivent à un rythme rapide et ont une distribution dynamique. L'exploration des flux de données est un domaine de recherche actif qui a récemment évolué afin de découvrir des connaissances à partir des vastes volumes de données générées en permanence. Cette thèse apportera des contributions dans les contextes de l'agriculture intelligente et des C-ITS, en mettant l'accent sur les C-ITS.

L'IoT dans l'agriculture fait référence à l'utilisation de capteurs et d'autres dispositifs pour convertir en données chaque aspect et opération impliqués dans l'agriculture. L'agriculture a grandement bénéficié des récentes avancées dans la technologie des capteurs, la science des données et les approches d'apprentissage automatique. Ces innovations sont une réponse aux défis environnementaux et démographiques auxquels notre société est confrontée, où de grandes augmentations de la production agricole mondiale sont nécessaires pour nourrir une population croissante. On prévoit que l'utilisation d'approches innovantes dans l'IdO contribuera à améliorer la production agricole de 70% d'ici 2050 [1]. Par conséquent, il est essentiel d'analyser les données générées afin de recueillir des informations susceptibles de faciliter la prise de décision et l'amélioration de la productivité.

Les systèmes de transport intelligents coopératifs (STIC) avec des véhicules connectés en réseau sont prêts à remodeler l'avenir de la mobilité. Cette transition est facilitée par les échanges de messages entre véhicules (V2V) et entre véhicules et infrastructures de transport (V2I). Les informations en temps réel sur les véhicules individuels proviennent des messages de sensibilisation coopératifs (CAM). Néanmoins, en raison de la nouveauté du concept, l'impact des services C-ITS sur les réseaux routiers n'a pas encore été complètement ressenti et évalué [2].

Les incidents de circulation sont des événements non récurrents qui peuvent provoquer des embouteillages et des retards dans les déplacements. Un incident est "un événement inattendu qui perturbe temporairement le flux de circulation sur un segment de route" [3]. Il est important de comprendre la fréquence des incidents en identifiant les variations par rapport aux schémas de circulation réguliers afin de réduire l'effet et la durée des incidents. Les accidents de véhicules, les pannes de véhicules, les débris sur la route et les véhicules arrêtés au milieu de la route sont tous des exemples d'incidents/anomalies routiers. Une détection précoce de ces anomalies permet de réduire les risques d'incidents tels que les accidents et les embouteillages. La plupart de ces incidents peuvent être attribués au comportement du conducteur et à l'état de la route. Les usagers de la route et les autorités ont tout à gagner à connaître le lieu, l'heure et la fréquence de ces anomalies routières.

L'objectif principal du C-ITS est d'améliorer la sécurité, le confort, le trafic et l'efficacité énergétique. La vitesse du véhicule et d'autres indicateurs basés sur la vitesse sont des paramètres couramment utilisés dans la recherche sur le trafic pour générer des profils de conduite [4], [5]. Le principal objectif de l'étude de la variation de la vitesse est de mieux comprendre pourquoi les conducteurs réagissent de telle ou telle manière aux conditions de la route et du trafic et de découvrir les facteurs qui influencent leurs actions. Les informations obtenues à partir de l'analyse des vitesses des véhicules peuvent être utiles pour identifier les points noirs (lieux propices aux accidents) et pour mieux comprendre la durée des trajets. En outre, elles peuvent être utilisées pour évaluer la nécessité de mettre en place des infrastructures, par exemple des ralentisseurs, des bosses et des radars sur certaines sections de la route.

Dans cette thèse, nous abordons le problème de l'analyse des données IoT en nous concentrant sur la détection des anomalies dans les flux de données et l'analyse du comportement des conducteurs. Nous considérons une approche orientée données avec l'objectif de détecter des anomalies à la volée en utilisant des

approches de détection non supervisées pour la détection d'anomalies contextuelles locales. Nous proposons une amélioration d'un détecteur d'anomalies d'ensemble, Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP). ELSCP est adapté au contexte du streaming en utilisant un cadre de pipeline qui met en œuvre une technique de fenêtre glissante. Cette adaptation permet le traitement des données sous forme de flux, ce qui est important dans la mesure où les performances de notre algorithme peuvent être évaluées dans le contexte des flux de données.

Il est essentiel pour la sécurité routière de comprendre les mécanismes de sélection de la vitesse des conducteurs et les variables de risque qui influent sur leur comportement en matière de vitesse. Nous avons proposé une approche de segmentation et d'analyse statistique des données pour l'analyse des signatures de vitesse continues générées dans des segments de route contigus. Nous considérons le mouvement collectif des véhicules le long d'un segment de route particulier et évaluons leur comportement global de conduite par l'analyse des signatures de vitesse.

L'exploration de trajectoires repose sur l'analyse des traces de mouvement / trajectoires pour l'extraction de modèles comportementaux. Pour caractériser les aspects comportementaux et le style de vie d'une entité, il est impératif d'analyser les traces quotidiennes. Un principe de confidentialité est appliqué dans le C-ITS où chaque message envoyé se voit attribuer un identifiant du véhicule émetteur. Cet identifiant n'est conservé que pendant un intervalle de temps donné, ce qui signifie qu'un véhicule peut avoir plusieurs identifiants. Pour reconstituer le mouvement du véhicule sur une longue période de temps, les identifiants des trajets consécutifs doivent être identifiés par un processus de liaison. Nous proposons de résoudre le problème de liaison des données en enchaînant des trajectoires anonymes à des véhicules potentiels en considérant la similarité des modèles de mouvement.

Contribution

Nous proposons une méthodologie de détection des anomalies dans les flux de données qui a été appliquée dans différentes études de cas présentées dans le chapitre 4. Notre première contribution sur la détection d'anomalies dans les flux de données en présence d'obstacles et de dérive de concept a été présentée à Nets4Workshop : 15th International Workshop on Communication Technologies for Vehicles [6].

Des travaux sur la détection des comportements anormaux des moissonneuses-batteuses à l'aide d'une technique d'ensemble avec mise en œuvre du concept de fenêtre glissante ont été présentés à la conférence International Conference on Smart and Sustainable Agriculture (SSA'2021) [7]. En outre, des travaux sur la détection d'anomalies pouvant être liées à l'état ou à la santé des cultures pendant la récolte et celles qui ont un impact sur l'efficacité de la récolte ont été publiés dans le journal Agriculture [8]. Des travaux sur la détection d'anomalies à partir de messages C-ITS utilisant un ensemble non supervisé et le concept de fenêtre glissante ont été soumis à IEEE Transactions on Intelligent Transportation Systems [9].

L'analyse des signatures de vitesse générées par les messages C-ITS a également été envisagée dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturaliste. Ce travail, développé dans le chapitre 5, a été présenté à la conférence IEEE International Conference on Communications (ICC 2021) [10].

Le dernier problème, celui de la liaison des trajectoires, est étudié en enchaînant les trajectoires des messages C-ITS à des véhicules potentiels en considérant la similarité des modèles de mouvement. Ce sujet est abordé dans le chapitre 6 et un article a été présenté à la Conférence 2020 IEEE Global Communications Conference (GLOBECOM 2020) [11].

En résumé, il y a un total de cinq publications acceptées et une soumise :

1. Juliet Chebet Moso, Ramzi Boutahala, Brice Leblanc, Hacène Fouchal, Cyril de Runz, Stéphane Cormier, and John Wandeto. Anomaly Detection on Roads Using C-ITS Messages. In *International Workshop on Communication Technologies for Vehicles*, pages 25–38. Springer, November 2020.
2. Juliet Chebet Moso, Stéphane Cormier, Cyril de Runz, Hacène Fouchal, and John Mwangi Wandeto. Anomaly Detection on Data Streams for Smart Agriculture. *Agriculture*, 11(11), 2021. ISSN 2077-0472. Doi: 10.3390/agriculture11111083.
3. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Mwangi Wandeto. Streaming-based Anomaly Detection in ITS Messages. Submitted to *IEEE Transactions on Intelligent Transportation Systems*, 2021.
4. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Mwangi Wandeto. Abnormal behavior detection in farming stream

- data. In: Boumerdassi S., Ghogho M., Renault É. (eds) *International Conference on Smart and Sustainable Agriculture. SSA 2021., Communications in Computer and Information Science*, vol 1470, pp. 44-56. Springer, Cham. June 2021.
5. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, John M. Wandeto, and Hasnaâ Aniss. Road Speed Signatures from C-ITS messages. In *ICC 2021-IEEE International Conference on Communications*, pages 1–6. IEEE, June 2021.
 6. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Wandeto. Trajectory User Linking in C-ITS Data Analysis. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, December 2020.

Plan de la thèse

Cette thèse est organisée comme suit :

Le chapitre 2 présente l'état de l'art. Il commence par décrire le domaine de l'agriculture intelligente dans la section 2.2 puis se poursuit par celui des systèmes de transport intelligents dans la section 2.3. La détection d'anomalies est présentée dans la section 2.4 où diverses approches et applications sont discutées. La section 2.5 présente les concepts et les applications de l'exploration des données de trajectoire et est suivie de la section 2.6 sur l'analyse du profil de vitesse des véhicules. Il se termine par la section 2.7 avec un examen des approches de liaison entre utilisateurs de trajectoires.

Le chapitre 3 présente les définitions et les méthodes. Il présente une discussion sur les mesures de performance et les mesures qui seront utilisées dans l'évaluation des performances des différents algorithmes et méthodologies utilisés. Une discussion sur la structure et les règles de génération de la CAM est présentée ainsi que les outils de simulation et la procédure utilisée pour générer des données CAM simulées.

Le chapitre 4 est consacré à la détection des anomalies. Nous proposons une technique ELSCP (Enhanced Locally Selective Combination in Parallel outlier ensembles). Nous définissons une méthodologie non supervisée basée sur les données et l'appliquons à trois études de cas : la détection des dommages aux cultures dans un ensemble de données sur les cultures, l'application aux journaux GPS des moissonneuses-batteuses et l'application aux messages C-ITS. L'objectif principal

est d'identifier les anomalies qui peuvent être liées à l'état ou à la santé des cultures pendant la récolte, celles qui ont un impact sur l'efficacité de la récolte et celles qui ont un impact sur la sécurité et l'efficacité des routes. Nous utilisons des techniques basées sur des tests d'hypothèses sur l'occurrence d'un modèle de mouvement anormal pendant la récolte dans le champ et sur les routes.

Le chapitre 5 traite de la génération et de l'analyse des signatures de vitesse à partir des données C-ITS par l'application de la segmentation des trajectoires et de l'analyse statistique. Les approches proposées dans ce chapitre utilisent un ensemble de données réelles de CAMs collectées dans le cadre du projet SCOOP@F [12] dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturaliste.

Le chapitre 6 considère le problème de la liaison trajectoire-utilisateur (TUL) et l'applique aux messages générés par les véhicules dans un environnement C-ITS. Le problème TUL est résolu par le chaînage de trajectoires anonymes à des véhicules potentiels en considérant la similarité des modèles de mouvement.

Le chapitre 7 conclura cette thèse et proposera quelques travaux futurs pour le travail présenté.

Synthèse

Dans cette synthèse, un résumé des informations importantes de chaque chapitre sera présenté.

Chapitre 2 : État de l'art

Notre état de l'art présente une discussion sur l'agriculture intelligente, les systèmes de transport intelligent (STI), les techniques de détection des anomalies et leur application dans les flux de données, les données agricoles et la gestion du trafic routier. Il aborde également l'exploration des trajectoires, l'analyse des profils de vitesse et les approches de liaison des trajectoires.

Agriculture intelligente

La population mondiale continue de croître et pourrait nécessiter une augmentation significative de la production alimentaire. La production agricole dépend de la gestion des sols, de l'eau, des ravageurs, des conditions climatiques et des risques imprévus, pour une production agricole durable [13]. L'utilisation de pesticides est un problème majeur ayant un impact sur l'environnement, la santé et la

sécurité des agriculteurs [14]. L'agriculture intelligente s'attaque aux difficultés de la production agricole en termes d'efficacité, d'impact environnemental, de sécurité alimentaire et de viabilité à long terme. L'IoT a le potentiel de fournir des solutions à une variété de préoccupations agricoles traditionnelles, notamment la réponse à la sécheresse, l'amélioration du rendement, l'irrigation et la gestion des pesticides.

Systèmes de transport intelligents

Les systèmes de transport intelligents offrent des avancées substantielles en matière de sécurité, de mobilité, de productivité et de durabilité environnementale des systèmes de transport. Les micro-systèmes de transport spécialisés sont constitués d'un nombre limité de véhicules coopérant à une tâche spécialisée (par exemple, la récolte où coopèrent moissonneuses-batteuses, charrettes à grains et semi-remorques) [15]. Une bonne coordination entre ces véhicules permet d'atteindre un niveau d'efficacité raisonnable dans l'exécution de la tâche.

L'objectif des C-ITS est la connectivité des véhicules afin que les conducteurs aient une bonne connaissance des conditions de circulation sur la route [16]. Les C-ITS communiquent à travers deux types de réseaux: le réseau WiFi avec le protocole 802.11p / G5 [17] qui est principalement basé sur des messages de diffusion et le réseau cellulaire. Les messages décentralisés de notification environnementale (DENM) contiennent des informations sur un événement ayant un impact potentiel sur la sécurité routière ou les conditions de circulation (par exemple, travaux routiers, accidents, pannes de véhicules, etc.).

Détection des anomalies

Les anomalies sont des "modèles dans les données qui ne se conforment pas à une notion bien définie de comportement normal" [18]. Une discussion est présentée sur les types d'anomalies, les techniques, les applications et les défis rencontrés dans la détection des anomalies (par exemple, la malédiction de la dimensionnalité, la sélection et l'ajustement des hyperparamètres). Le type d'anomalies qui nous intéresse dans cette thèse sont les anomalies contextuelles.

Plusieurs approches applicables à la détection d'anomalies dans les flux de données ont été discutées. STORM [19] est un algorithme de distance basé sur le modèle de la fenêtre glissante. DBOD-DS [20] est basé sur une fonction de densité de probabilité adaptative. La théorie des valeurs extrêmes est adoptée par SPOT pour les flux stationnaires et DSPOT pour les flux avec dérive du concept [21],

[22]. Dans l'environnement de l'Internet des véhicules, SafeDrive utilise un modèle de graphe d'état pour l'identification des comportements de conduite anormaux. Notre approche applique le concept de fenêtre glissante dans la détection des anomalies.

La détection de conditions anormales dans l'environnement agricole a vu l'application de la régression linéaire aux données des capteurs[23] où les données anormales peuvent indiquer que l'environnement agricole n'est pas propice à la croissance des cultures. L'apprentissage profond a également été appliqué à la détection d'obstacles et d'anomalies dans un champ agricole à l'aide de Deep-Anomaly [24], à la détection de l'occurrence de catastrophes régionales dues au gel à l'aide de SVM et d'ANNs [25] et à la détection de lectures anormales de données de capteurs à l'aide de modèles LSTM et ARIMA [26]. Le traitement d'images a également été appliqué à la détection de la croissance aberrante de la végétation en utilisant la forêt d'isolement [27]. Un filtre de Kalman et l'algorithme DBSCAN ont également été utilisés pour détecter les mouvements anormaux des moissonneuses-batteuses [28].

La détection des anomalies dans les données du trafic routier peut aider à découvrir des modèles inhabituels dans le trafic, qui peuvent être analysés plus en détail pour révéler des incidents de circulation. La congestion du trafic et la gestion des routes sont deux types d'anomalies du trafic [29]. Pour les anomalies locales du trafic, le réseau routier est divisé en segments indépendants, puis les anomalies individuelles sont extraites par segment. Dans la catégorie des anomalies de groupe, une anomalie dans un segment de route se propage à d'autres routes adjacentes et est analysée en considérant les interactions causales entre les segments de route.

Extraction de données sur les trajectoires

L'exploration de trajectoires peut être considérée comme un processus d'analyse des traces de mobilité dans le but de découvrir des modèles spatiaux, spatio-temporels et comportementaux par le biais du regroupement, de la classification, de la détection d'anomalies et de la détection de lieux intéressants [30]. Le défi actuel consiste à exploiter ces données pour en extraire des connaissances et des informations utiles à l'amélioration des niveaux de mobilité [31]. Une caractéristique discriminante commune est la distance entre deux trajectoires ou segments de sous-trajectoire qui est calculée à l'aide d'une mesure ou d'une métrique de

distance basée sur le type d'application. C'est l'approche que nous appliquons dans le calcul de la similarité des modèles de mouvement.

Analyse du profil de vitesse

La vitesse du véhicule et d'autres indicateurs basés sur la vitesse sont des paramètres couramment utilisés dans la recherche sur le trafic pour générer des profils de conduite. L'objectif principal de l'analyse des profils de vitesse est de mieux comprendre pourquoi les conducteurs réagissent comme ils le font à différentes conditions de circulation ou de route, et d'identifier les facteurs qui influencent leur comportement. L'analyse des signaux de vitesse et d'accélération a trouvé des applications dans la détection des feux de signalisation, des croisements de rues, des ronds-points [32], des panneaux d'arrêt [33]. [34] et la génération de cartes routières de haute qualité [35]. En outre, il est possible de détecter en temps réel les encombrements et les incidents de la circulation routière à l'aide des données sur la vitesse des véhicules [36].

Liaison Trajectoire-Utilisateur (TUL)

L'établissement d'un lien entre la trajectoire et l'utilisateur est motivé par le fait que les applications LBSN génèrent un grand nombre de données qui sont généralement dépouillées des identifiants de l'utilisateur afin d'anonymiser les données et de préserver la vie privée [37]. Les approches d'apprentissage profond sont utilisées dans l'état de l'art pour traiter les modèles de mobilité sémantique, l'hétérogénéité et la rareté des données de mobilité lors de la résolution de TUL. Lier ces trajectoires aux utilisateurs qui les ont générées peut fournir des informations précieuses pour les systèmes de recommandation et l'identification de criminels par le biais de signaux téléphoniques et de check-ins, entre autres applications.

Chapitre 3 : Contexte et méthodes

Ce chapitre est consacré aux définitions des termes communs utilisés dans les trois chapitres suivants. Il présente également une discussion sur les paramètres et les mesures de performance qui seront utilisés dans l'évaluation des performances des différents algorithmes et méthodologies utilisés. Les indicateurs de performance sont l'aire sous la courbe de la caractéristique opérationnelle du récepteur (AUC-ROC), l'aire sous la courbe de la précision du rappel (AUCPR) et le score F1. Il est essentiel d'évaluer avec précision les similitudes des différentes approches afin de choisir la meilleure. Une discussion sur la structure et les règles de génération

de la CAM est présentée ainsi que les outils et la procédure utilisée pour générer des données CAM simulées réalistes.

La simulation d'applications de réseaux ad hoc véhiculaires (VANET) nécessite la simulation de la communication sans fil entre les véhicules et de la mobilité de ces derniers. Les simulations ont été effectuées à l'aide du simulateur de réseau OMNET++ [38], et du simulateur de trafic routier SUMO [39]. Artery V2X Simulation Framework [40] est utilisé pour intégrer le simulateur de réseau et le simulateur de trafic routier, ce qui facilite la communication. Cette intégration est particulièrement importante pour le C-ITS qui met en œuvre des applications de sécurité et d'efficacité du trafic. La simulation a porté sur la mobilité des véhicules équipés de C-ITS dans la ville de Reims, en France. La mobilité et les communications entre les véhicules ont été réalisées de manière à imiter le mouvement réel dans l'environnement C-ITS. Ceci afin de garantir que les résultats produits soient aussi réalistes que possible.

Chapitre 4 : Algorithme LSCP amélioré pour la détection des anomalies dans les données IoT

Nous considérons une approche orientée données avec l'objectif de détecter des anomalies à la volée en utilisant des approches de détection non supervisées pour la détection d'anomalies contextuelles locales. Nous proposons un détecteur d'anomalies par ensemble appelé Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP). ELSCP est adapté au contexte du streaming en utilisant un système qui convertit les données en un flux et les transmet à ELSCP en utilisant un modèle de fenêtre de référence qui met en œuvre une technique de fenêtre glissante. Cette adaptation permet le traitement des données sous forme de flux, ce qui facilite l'évaluation de notre algorithme dans le contexte du streaming.

Nous étudions la détection d'anomalies sur trois axes: la détection des dommages causés aux cultures pendant la récolte, l'utilisation efficace des moissonneuses-batteuses et la détection de situations anormales sur la route. Pour chaque cas, nous utilisons un ensemble de données pertinentes pour identifier les anomalies. Nous proposons des techniques basées sur des tests d'hypothèses sur l'occurrence d'un modèle de mouvement anormal pendant la récolte dans le champ et sur les routes. L'hypothèse principale de notre analyse est que "les instances normales sont beaucoup plus fréquentes que les anomalies". L'hypothèse clé est que "des points similaires dans un espace caractéristique ont des scores d'anomalie similaires".

Dans l'ensemble de données sur les cultures, notre analyse a montré que 30% des anomalies détectées pouvaient être directement liées aux dommages causés aux cultures. Dans le jeu de données sur les moissonneuses-batteuses, notre approche obtient les meilleurs scores avec une AUC-ROC supérieure de 6,4% à celle de la deuxième approche COPOD (99,8% contre 93,4%); une AUCPR de 97,2% alors que les meilleurs scores des autres approches (celui de l'OCSVM) ne sont que de 38,5%. Il convient également de mentionner que notre méthodologie permet de détecter efficacement les comportements déviants des moissonneuses-batteuses. Par conséquent, la détection des anomalies pourrait être intégrée dans le processus de décision des exploitants agricoles afin d'améliorer l'efficacité de la récolte et la santé des cultures.

Dans le contexte du C-ITS, nous avons considéré les anomalies qui pourraient avoir des conséquences telles qu'un accident/incident ou un véhicule immobilisé sur la route qui oblige les véhicules sur cette section particulière de la route à réduire considérablement leur vitesse lorsqu'ils rencontrent le point d'incident. Nous constatons que ELSCP surpasse toutes les autres techniques, avec des performances en AUC-ROC supérieures de 3,64% (89,45% contre 85,81%) et en AUCPR supérieures de 9,83% (38,41% contre 28,58%) à celles de la deuxième meilleure technique LSCP. Cette détection pourrait être intégrée dans le processus de décision des opérateurs routiers afin d'améliorer la sécurité et la fluidité du trafic.

Chapitre 5 : Signatures de données à partir de données IoT

Il est essentiel pour la sécurité routière de comprendre les mécanismes de sélection de la vitesse des conducteurs et les variables de risque qui influent sur leur comportement en matière de vitesse. Nous avons envisagé l'analyse des signatures de vitesse générées par les messages C-ITS dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturaliste. Les données C-ITS sont utilisées pour évaluer la pertinence des limites de vitesse indiquées sur l'itinéraire considéré.

La génération et l'interprétation des signatures de vitesse sont réalisées en évaluant les données à l'aide de techniques d'analyse statistique descriptive. Sur la base de l'analyse de la variation de la vitesse médiane, nous pouvons conclure que pour la plupart des parties de la route, en amont comme en aval, la politique de limitation de vitesse est respectée.

En ce qui concerne la plage de vitesse préférée par conducteur, les résultats montrent que sur la section amont, à l'exception de deux conducteurs, tous les

autres conducteurs respectent la limite de vitesse dans 80% des cas. La plage de vitesse la plus préférée est de 11 à 80 km/h. Sur la section aval, à l'exception de deux conducteurs, tous les autres conducteurs ont dépassé la limite de vitesse à un moment donné. La majorité des conducteurs dépassant la limite de vitesse, il pourrait être nécessaire de revoir la limite de vitesse ou d'introduire une limite de vitesse variable afin d'atténuer le comportement affiché.

Une application possible des signatures de vitesse générées est celle des véhicules autonomes, où elles peuvent être utilisées comme données d'entrée afin de vérifier si la conduite respecte la vitesse normale ou le comportement naturel attendu. Les signatures peuvent également être utilisées par les ingénieurs de la circulation et les planificateurs des transports pour évaluer l'efficacité des limites de vitesse affichées, de sorte que des ajustements puissent être effectués en fonction du comportement observé des conducteurs. Cela peut nécessiter une réduction ou une augmentation de la limite de vitesse actuelle en évaluant la vitesse du 85e percentile du flux de données sur le trafic.

Chapitre 6 : Liaison des données IoT

Les véhicules d'un réseau de transport intelligent échangent un grand nombre de messages. Chaque message envoyé est généré avec un identifiant du véhicule émetteur. Pour respecter la vie privée des utilisateurs, un identifiant n'est conservé que pendant un intervalle de temps déterminé. Le besoin qui se pose est le suivant : étant donné que plusieurs identifiants sont attribués à un véhicule, sommes-nous capables de regrouper les identifiants et de détecter ceux qui appartiennent au même véhicule ? Nous avons résolu ce problème en enchaînant des trajectoires anonymes à des véhicules potentiels en considérant la similarité des schémas de mouvement.

Notre objectif était de relier les identifiants qui se sont produits à la même date dans un court laps de temps (quelques secondes) afin d'obtenir des trajectoires continues dans l'espace et le temps. En outre, pour tester la continuité, les trajectoires correspondantes devaient se déplacer dans la même direction lors du changement d'identifiant. Le fait que les trajectoires soient contraintes par un réseau routier augmente la probabilité de relier les segments de trajectoire à l'utilisateur générateur.

Sur les 3866 trajectoires, nous n'avons pu relier que 867 paires, soit 22,43% du nombre total de trajectoires. D'après notre analyse, il est possible de relier les trajectoires aux utilisateurs générateurs si d'autres attributs distinctifs (comme la

vitesse, l'angle de cap, l'altitude et la direction de conduite) et les connaissances de base sur la génération des messages sont pris en compte lors de l'analyse de similarité. Cependant, l'association complète de tous les segments aux utilisateurs générateurs est une tâche difficile et pourrait ne pas être possible. Il convient également de noter que l'utilisation de pseudonymes comme mesure de confidentialité et de sécurité s'est avérée être une approche viable puisque nous n'avons pas été en mesure de briser l'exigence de non-liaison.

Conclusion et perspectives

Conclusion

Dans cette thèse, nous avons abordé le problème de l'analyse des données IoT en mettant l'accent sur la détection des anomalies dans les flux de données et l'analyse du comportement des conducteurs. Un examen de l'état de l'art a été réalisé dans le chapitre 2 où une discussion détaillée sur l'agriculture intelligente, les STI, la détection d'anomalies, l'exploration de données de trajectoire, l'analyse du profil de vitesse et les approches de liaison entre utilisateurs de trajectoire a été présentée. Nous avons apporté des contributions dans les contextes de l'agriculture intelligente et des C-ITS, en mettant l'accent sur les C-ITS.

De nombreux développements ont été réalisés dans le domaine des systèmes de détection des anomalies, avec de nombreuses techniques proposées pour résoudre le problème de l'identification des anomalies. Les domaines d'application de la détection d'anomalies sont également très variés, ce qui nécessite une solution fiable et précise. L'apprentissage non supervisé est privilégié pour les applications de la vie réelle, en particulier pour la détection d'anomalies, car il y a beaucoup de données sans étiquettes dans ce scénario. Dans le chapitre 4, nous avons proposé un nouvel algorithme basé sur un détecteur d'anomalies d'ensemble appelé Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP) pour la détection d'anomalies de flux. Sur cette base, nous avons défini une méthodologie non supervisée pilotée par les données qui est appliquée dans trois études de cas, avec l'objectif de détecter les anomalies à la volée.

Dans le premier cas d'étude, nous avons effectué une détection non supervisée d'anomalies sur des données de cultures. Nous avons étudié le lien entre l'état des cultures (endommagées ou non) et les anomalies détectées en analysant les données sur les dommages aux cultures enregistrées par les agriculteurs sur une période de trois saisons de récolte. Sur la base de nos résultats, il a été possible de relier les

anomalies extraites de l'analyse multivariée de diverses caractéristiques à l'état de la culture endommagée à la fin de la récolte. La deuxième étude est consacrée à la détection d'anomalies dans les flux de données obtenus à partir des traces GPS des moissonneuses-batteuses collectées pendant la récolte du blé. D'après nos résultats, notre méthodologie permet de détecter efficacement les comportements déviants des moissonneuses-batteuses. Les performances obtenues ont été évaluées par rapport à d'autres approches, et les résultats obtenus sont pertinents (sur la base de métriques de performance connues) avec la possibilité de sa mise en œuvre en temps réel pour détecter les anomalies et aider les agriculteurs pendant la récolte. Par conséquent, la détection des anomalies pourrait être intégrée dans le processus de décision des exploitants agricoles afin d'améliorer l'efficacité de la récolte et la santé des cultures.

Dans le troisième cas d'étude, les messages CAM C-ITS recueillis dans le cadre d'un projet de collaboration entre des opérateurs routiers, des constructeurs automobiles et des universitaires ont été analysés afin de détecter les anomalies sur la route. Nous considérons les anomalies qui pourraient avoir des conséquences telles qu'un accident/incident ou un véhicule bloqué sur la route qui oblige les véhicules sur cette section particulière de la route à réduire considérablement leur vitesse lorsqu'ils rencontrent le point d'incident. Nous avons utilisé des approches en continu car, dans les environnements C-ITS réels, la détection des anomalies serait mise en œuvre dans les unités routières qui recueillent de nombreux messages des véhicules à portée. Cela signifie qu'il faudrait traiter les messages à la volée pour de nombreuses raisons (limitation de la mémoire, temps de réponse, etc.). D'après nos résultats, ELSCP est capable de détecter les anomalies des CAMs. Cette détection pourrait être intégrée dans le processus de décision des opérateurs routiers afin d'améliorer la sécurité et la fluidité du trafic. La détection rapide des anomalies est essentielle, notamment pour les équipes d'intervention d'urgence, ce qui permet d'améliorer l'efficacité des opérations de sauvetage.

Dans le chapitre 5, nous avons considéré l'analyse des signatures de vitesse générées à partir des messages C-ITS dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturaliste. Nous avons montré qu'avec l'application de la segmentation et des statistiques agrégées, il est possible d'obtenir une meilleure compréhension du comportement général de conduite et de déduire des informations relatives à l'état de la route et à la situation du trafic. Avec l'adoption actuelle du C-ITS, il reste un problème de quantité insuffisante de données pour une analyse plus détaillée des profils de conduite à un niveau microscopique (niveau de l'itinéraire).

Dans le chapitre 6, nous avons considéré le problème de liaison de trajectoire et l'avons appliqué aux messages générés par les véhicules dans le C-ITS. D'après notre analyse, il est possible de relier les trajectoires aux utilisateurs qui les ont générées si d'autres attributs distinctifs (comme la vitesse, l'angle de cap, l'altitude et la direction de conduite) et les connaissances de base sur la génération des messages sont pris en compte lors de l'analyse de similarité. Il est également intéressant de noter que l'utilisation de pseudonymes comme mesure de confidentialité et de sécurité s'est avérée être une approche viable puisque nous n'avons pas été en mesure de briser l'exigence de non-liaison.

Limites

La première limitation d'ELSCP concerne l'extraction des points de données voisins qui constituent la région locale d'une instance de test en utilisant des mesures de distance appliquées à l'algorithme KNN Ball Tree. Cette approche pose deux problèmes :

1. Il faut beaucoup de temps pour déterminer les plus proches voisins de l'instance de test ;
2. Les performances dans un espace multidimensionnel peuvent être affectées, notamment lorsque de nombreuses caractéristiques ou attributs ne sont pas pertinents.

Pour remédier à ce problème, la définition de la région locale pourrait être résolue par l'utilisation de méthodes approximatives rapides [41] ou par le prototypage [42], ce qui peut réduire de manière significative le temps requis pour établir le domaine local car tous les points de données ne sont pas requis par ces techniques. La deuxième limite concerne la génération de la pseudo-vérité terrain, pour laquelle nous avons appliqué une technique de maximisation simple. Cela pourrait être amélioré en considérant des stratégies exactes, par exemple, avec l'élagage actif des détecteurs de base [43].

La troisième limite concerne la disponibilité des données du C-ITS. Avec l'adoption actuelle du C-ITS, il reste un problème de quantité insuffisante de données pour une analyse plus détaillée des profils de conduite à un niveau microscopique (niveau de la route). Avec les données actuelles, il n'est pas possible d'analyser efficacement les données pour une analyse horaire ou hebdomadaire du comportement de conduite sur des routes spécifiques.

Perspectives

Nos travaux futurs se concentreront sur la calibration des scores aberrants en introduisant des fonctions de perte dépendantes, car un faux négatif dans les scénarios d'agriculture intelligente et de C-ITS peut induire des problèmes gênants, notamment sur la production agricole, l'efficacité des exploitations et la sécurité routière.

Nous proposons également d'introduire un calibrage automatique des paramètres dans le but d'améliorer les chances de déploiement de l'algorithme pour les opérateurs d'infrastructures agricoles et routières. Une autre partie critique sera l'optimisation du processus afin de gagner en complexité avec des ajustements fins des règles de décision de l'apprentissage d'ensemble pour améliorer l'efficacité.

La configuration des sous-espaces locaux peut également être améliorée pour diminuer le temps passé à localiser les plus proches voisins d'une instance de test en remplaçant la stratégie de recherche kNN par une technique de clustering. Nous proposons également de développer la méthodologie ELSCP en un outil pouvant être utilisé pour la détection et l'analyse en temps réel de flux de données.

Dans l'analyse des signatures de vitesse et des profils de conduite, il sera intéressant de faire une analyse plus détaillée des données pour identifier les facteurs causant ou influençant le comportement observé, en particulier sur les points de pointe. Les messages DENM peuvent également être utilisés pour extraire les points où des incidents ont été signalés, puis utiliser ces emplacements comme POI pour la détection des incidents de circulation. En comparant les signatures de vitesse extraites des segments de route avec les points d'incidents connus provenant des DENM, il pourrait maintenant être possible de mieux comprendre l'évolution du comportement de conduite.

Pour résoudre le problème de l'insuffisance des données dans le C-ITS, les signatures de vitesse générées dans un environnement de conduite naturaliste réel dans cette étude peuvent être utilisées pour générer des données synthétiques à partir des caractéristiques apprises des modèles de mouvement des véhicules. Les signatures peuvent également être appliquées dans les véhicules autonomes où les signatures peuvent être utilisées comme entrées afin de vérifier si la conduite suit le comportement de vitesse naturaliste attendu.

Il existe un certain nombre de stratégies de changement de pseudonyme dans le C-ITS qui affectent la manière dont les identifiants sont changés. Les données utilisées dans cette étude appliquaient une stratégie *round robin*. Il serait intéressant d'évaluer la méthodologie de liaison de données développée dans cette thèse

sur des données provenant d'autres stratégies de changement afin de valider son efficacité.

Contents

Abstract	iii
Résumé court	iv
Acknowledgements	v
French Summary	ix
List of Figures	xxxi
List of Tables	xxxv
Abbreviations	xxxvii
1 Introduction	1
1.1 Context	1
1.2 Contribution	3
1.3 Outline	5
2 State of the art	7
2.1 Introduction	7
2.2 Smart Agriculture	7
2.3 Intelligent Transport Systems	9
2.3.1 Specialized Micro Transportation Systems	9
2.3.2 Cooperative Intelligent Transport Systems	10
2.4 Anomaly detection	12
2.4.1 Anomaly types and detection techniques	13
2.4.1.1 Statistical approaches	15
2.4.1.2 Ensemble-based approaches	18
2.4.1.3 Density-based approaches	20
2.4.1.4 Copula-based approach	22
2.4.1.5 Linear-based approach	23
2.4.2 Anomaly detection in Data streams	23
2.4.3 Anomaly detection in agricultural data	25
2.4.4 Road traffic anomaly detection	27

2.5	Trajectory data mining	30
2.6	Speed profile analysis	33
2.7	Trajectory-User Linking (TUL)	35
2.8	Discussion	36
3	Background and methods	39
3.1	Introduction	39
3.2	Definitions	39
3.3	Performance indicators	40
3.4	CAM data generation	42
3.4.1	Simulation software	45
3.4.2	Data generation	46
4	Enhanced LSCP Algorithm for anomaly detection on IoT data	49
4.1	Introduction	49
4.2	Problem statement	52
4.3	Proposed Enhanced LSCP Algorithm(ELSCP)	53
4.3.1	ELSCP Algorithm using Kendall rank correlation (ELSCP_K)	58
4.3.2	ELSCP Algorithm using Pearson correlation (ELSCP_P)	60
4.4	Application Scenarios	61
4.4.1	Scenario A: Crop dataset	62
4.4.2	Scenario B: Combine harvester GPS logs	63
4.4.3	Scenario C: C-ITS messages	66
4.5	Experimental Evaluation and Results	69
4.5.1	Scenario A: Crop damage	69
4.5.2	Scenario B: Combine harvester GPS data	72
4.5.3	Scenario C: C-ITS CAM data	74
4.6	Discussion	78
5	Data Signatures from IoT data	81
5.1	Introduction	81
5.2	Problem Statement and Methodology	83
5.2.1	Problem Statement	83
5.2.2	Methodology	84
5.3	Experimental Evaluation and Results	88
5.3.1	Evaluation of collective speed signatures	89
5.3.2	Evaluation of speed dispersion and preferred speed ranges	90
5.4	Discussion	97
6	IoT data linking	101
6.1	Introduction	101
6.2	Problem Statement and Methodology	103
6.2.1	Problem Statement	103
6.2.2	Methodology	103
6.2.3	Dataset Description	106

6.3	Experimental Evaluation and Results	107
6.4	Discussion	110
7	Conclusion	113
7.1	Limitations	115
7.2	Perspectives	115
	 Bibliography	 117

List of Figures

2.1	Representation of key technologies used in Smart Agriculture [44]	8
2.2	Representation of applications, services, and sensors used in Smart Agriculture [44]	9
2.3	Representation of V2V communication	12
2.4	Representation of V2I communication	13
2.5	Illustration of a point anomaly detection	14
2.6	Illustration of Contextual anomaly detection [45]	15
2.7	Illustration of Collective anomaly detection [46]	15
2.8	Anomaly detection workflow and anomaly detection techniques [47]	16
2.9	HBOS workflow for anomaly detection	16
2.10	LSCP flow chart. Steps requiring re-computation are highlighted in yellow; cached steps are in gray [48]	20
2.11	The reachability distance for different data points p with regard to o , when k equals 5 [49]	21
2.12	Movement trajectory of a tractor in a citrus grove [50]	28
3.1	CAM structure (ETSI EN 302 637-2) [51]	43
3.2	CAM generation rules (ETSI EN 302 637-2) [52]	44
3.3	Artery	46
3.4	Artery Architecture [40]	47
4.1	ELSCP flow chart: The results for steps 1 and 2 are cached; steps 3 - 5 are re-computation for each test instance	58
4.2	ELSCP workflow	61
4.3	Anomaly Detection framework	61
4.4	Histograms showing data distribution per attribute in the crop dataset	63
4.5	Heat map showing the correlation between the attributes of the crop dataset	64
4.6	Attribute-based Box plots showing the presence of outliers in the crop dataset	64
4.7	Field of interest: Trajectory of a combine harvester showing normal points in green and anomalies in red	66
4.8	Attribute-based Box plots showing the presence of outliers in combine harvester GPS logs	66
4.9	Heat map showing the correlation between the attributes of the combine harvester GPS logs dataset	67
4.10	Illustration of an accident scenario on the road	67

4.11	Area of interest; (a) trajectories on Boulevard Dauphinot (route N51), (b) Obstacle section: Normal instances in green, Anomalies in red	69
4.12	ELSCP ROC Curve indicating the best prediction threshold for crop dataset	71
4.13	ELSCP Precision-Recall Curve indicating the best prediction threshold for crop dataset	71
4.14	ELSCP ROC Curves performance evaluation for Combine harvester GPS dataset	73
4.15	ELSCP Precision-Recall performance evaluation Combine harvester GPS dataset	73
4.16	Comparison of models' AUC-ROC performance	75
4.17	Comparison of models' AUCPR performance	76
4.18	Comparison of models' Execution time performance	76
4.19	ELSCP ROC Curves performance evaluation	77
4.20	ELSCP Precision-Recall performance evaluation	77
5.1	Area of interest; (a) trajectories on the full stretch of route N118, (b) Entry point, (c) Exit point	86
5.2	Sampled road section showing the extracted one kilometre length segments	86
5.3	Upstream Speed signature variation.	91
5.4	Downstream speed signature variation.	91
5.5	Median speed signature variation for both upstream and downstream sections of the road.	93
5.6	Representation of road topology for segment 8 and 16 with upstream trajectory points in blue and downstream points in purple	93
5.7	Upstream Box-and-whisker plots of the drivers' vehicle speed ranges.	94
5.8	Downstream Box-and-whisker plots of the drivers' vehicle speed ranges.	95
5.9	Upstream Box-and-whisker plots of the vehicles' speed ranges per road segment.	96
5.10	Downstream Box-and-whisker plots of the vehicles' speed ranges per road segment.	96
5.11	Upstream Stacked bar chart of the drivers' chosen range of vehicle speeds.	97
5.12	Downstream Stacked bar chart of the drivers' chosen range of vehicle speeds.	98
6.1	Representation of movement of a C-ITS vehicle within 100 milliseconds based on a speed of 80 km/h	105
6.2	Trajectory mining framework.	106
6.3	Illustration of the pseudonym change strategy in the SCOOP@F project [53]	107
6.4	Distribution of all trajectories.	108

6.5	Distribution of origin-destination pairs.	109
6.6	Distribution of trajectories for the 5th and 6th of April 2019.	110
6.7	Continuity validation of linked trajectories.	110
6.8	Continuous trajectory after linking four trajectories.	111
6.9	Monthly analysis of linked trajectories.	111

List of Tables

3.1	Simulation parameters	48
4.1	Crop Data Description.	62
4.2	Performance comparison of various detectors on Crop dataset	72
4.3	Performance comparison of various detectors on combine harvester GPS Logs	74
4.4	Experimental parameters	74
4.5	ELSCP AUC-ROC and AUCPR performance results for window size variation	75
4.6	Average AUC-ROC and AUCPR performance results for various anomaly detection models on C-ITS data	78
5.1	Summary of vehicle speed	89
5.2	Upstream summary speed characteristics per road segment	90
5.3	Downstream summary speed characteristics per road segment	92
5.4	Upstream Drivers' speed box plot parameter values	94
5.5	Downstream Drivers' speed box plot parameter values	95

Abbreviations

ANNs	Artificial neural networks
AUC-ROC	Area under the curve of the receiver operating characteristic
AUCPR	Area Under the Curve of Precision-Recall
ARIMA	Autoregressive integrated moving average model
CAM	Cooperative Awareness Messages
CBLOF	Clustering-based local outlier factor
C-ITS	Cooperative Intelligent Transport Systems
CVTI	Common Visit Time Interval
COPOD	Copula-based outlier detector
DENM	Decentralized Environmental Notification Message
DBSCAN	Density-based spatial clustering of applications with noise
DBOD-DS	Distance-Based Outlier Detection for Data Streams
DSPOT	Drift Streaming Peak Over Threshold
DSRC	Dedicated Short Range Communication
EDR	Edit Distance on Real Sequence
ELSCP	Enhanced locally selective combination in parallel outlier ensembles
ETSI	European Telecommunications Standards Institute
FDM	Filter Discovery Match
FP	False positive
FPR	False Positive rate
FN	False negative
GPD	Generalized Pareto Distribution
GPS	Global positioning system
GPU	Graphics processing unit

HBOS	Histogram-based outlier score
IForest	Isolation Forest
IoT	Internet-of-Things
IQR	Interquartile range
ITS	Intelligent Transport Systems
ITS-C	ITS Central Station
kNN	k-nearest neighbours detector
LBSN	Location based social network applications
LCA	Lane Changing Advisor
LCSS	Longest common subsequence
LOF	Local outlier factor
LODA	Lightweight online detector of anomalies
LOF	Local Outlier Factor
LoTAD	Long-term traffic anomaly detection
LSTM	Long short-term memory
LSCP	Locally selective combination in parallel outlier ensembles
MAD	Median Absolute Deviation
MCD	Minimum covariance determinant
MSM	Multidimensional Similarity Measure
MSTP	Maximal Semantic Trajectory Pattern
OBU	On Board Unit
OCSVM	One-class support vector machines
P	Precision
POT	Peak Over Threshold
PyOD	Python outlier detection
QGIS	Quantum Geographic Information System
R	Recall
RNN	Recurrent Neural Networks
RSU	Road Side Unit
SAR	Synthetic Aperture Radar
SMSM	Stops and Moves Similarity Measure
SMTS	Specialized Micro Transportation Systems

SPOT	Streaming Peak Over Threshold
STORM	STream OutlieR Miner
SVM	Support vector Machine
TN	True negative
TP	True positive
TPR	True positive rate
TUL	Trajectory-User Linking
TULER	TUL via Embedding and RNN
TULVAE	TUL via Variational AutoEncoder
V2I	Vehicle-to-Infrastructure communication
V2V	Vehicle-to-Vehicle communication

*Dedicated to my loving parents Francis and Esther, my
loving husband Jonah and my adorable children Jeremy,
Jared and Jemimah. . .*

Chapter 1

Introduction

1.1 Context

Internet-of-Things (IoT) automates the execution of data-driven intelligent actions on connected devices by leveraging essential technologies such as sensor-based autonomous data collection and cloud-based big data analysis. This automation enables a wide range of practical real-world applications, including smart transportation, smart agriculture, smart cities, and so on. The incorporation of IoT into everyday life, has resulted in the collection of massive amounts of data. Data streams are vast, continuous, unbounded sequences of data that arrive at a fast rate and have a dynamic distribution. Data stream mining is an active research field that has recently evolved in order to uncover knowledge from the vast volumes of continually generated data. This thesis will make contributions in the contexts of smart agriculture and C-ITS, with a focus on C-ITS.

IoT in agriculture refers to the use of sensors and other devices to convert every aspect and operation involved in farming into data. Agriculture has benefited greatly from recent advances in sensor technology, data science, and machine learning approaches. These innovations are in response to the environmental and population challenges that our society is facing where large worldwide agriculture production increases are required to feed a growing population. It is anticipated that the use of innovative approaches in IoT will contribute to improving agricultural output by 70% by 2050 [1]. As a result, it is vital to analyze generated data in order to gather information that may aid in decision making and productivity enhancement.

Cooperative Intelligent Transport Systems (C-ITS) with networked vehicles are primed to reshape the future of mobility. This transition is facilitated by the flow

of messages between vehicles (V2V) and between vehicles and transport infrastructure (V2I). Real-time information on individual vehicles is derived from Cooperative Awareness Messages (CAM). Nonetheless, owing to the concept's newness, the impact of C-ITS services on road networks has yet to be completely felt and evaluated [2].

Traffic incidents are non-recurring events that may cause traffic congestion and travel time delays. An incident is "an unexpected event that temporarily disrupts the flow of traffic on a segment of a roadway" [3]. It is important to understand the frequency of incidents by identifying variations from regular traffic patterns to reduce the effect and length of incidents. Vehicle crashes, vehicle breakdowns, debris on the road, and vehicle(s) stopped in the middle of the road are all examples of road incidents / anomalies. An early detection of such anomalies will reduce incident risks such as accidents and traffic jam. Most of these incidents can be traced back to driver behaviour and road conditions. Road users and authorities benefit from knowing the place, time, and frequency of these road anomalies.

The main aim of C-ITS is to improve safety, comfort, traffic and energy efficiency. Vehicle speed and other speed based indicators are commonly used parameters in traffic research for generation of driving profiles [4], [5]. The main goal for studying speed variation is to gain a better understanding on why drivers respond in certain ways to road/traffic conditions and to discover factors which affect their actions. The information acquired from analysis of vehicle speeds can be useful in identification of black-spots (accident prone locations) and for gaining a better understanding of travel time. Further, it can be used to evaluate the need for infrastructure, for example speed bumps, humps and speed cameras on certain sections of the road.

In this thesis, we address the problem of analyzing IoT data with a focus on anomaly detection in data streams and driver behaviour analysis. We consider a data driven approach with the objective of detecting anomalies on the fly using unsupervised detection approaches for the detection of local contextual anomalies. We propose an enhancement of an ensemble anomaly detector, Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP). ELSCP is adapted to the streaming context using a pipeline framework which implements a sliding window technique. This adaptation enables the processing of data as a stream which is important in that the performance of our algorithm can be evaluated in the streaming context.

Understanding drivers' speed selection mechanisms and the risk variables impacting their speeding behaviour is critical for road safety. We proposed a segmentation and statistical data analysis approach for analysis of continuous speed signatures generated within contiguous road segments. We consider the collective movement of vehicles along a particular road segment and evaluate their aggregate driving behaviour through analysis of speed signatures.

Trajectory mining entails the analysis of movement traces / trajectories for extraction of behavioural patterns. To characterize the behavioural and lifestyle aspects of an entity, an analysis of daily traces is imperative. A privacy principle is applied in C-ITS where every message sent is assigned an identifier of the transmitting vehicle. This identifier is kept only over a specified time interval thus one vehicle will have multiple identifiers. To reconstruct the movement of the vehicle over a long period of time, the identifiers from the consecutive trips must be identified through a linking process. We propose to solve the data linking problem by chaining anonymous trajectories to potential vehicles by considering similarity in movement patterns.

1.2 Contribution

We propose a methodology for detection of anomalies in data streams which has been applied in different case studies discussed in chapter 4. Our first contribution on anomaly detection from data streams in presence of obstacles and concept drift was presented in Nets4Workshop: 15th International Workshop on Communication Technologies for Vehicles [6]. Work on abnormal behaviour detection of combine harvesters using an ensemble technique with implementation of sliding window concept was presented in International Conference on Smart and Sustainable Agriculture (SSA'2021) [7]. Additionally, work on detection of anomalies that can be linked to crop state/health during harvest and those that have an impact on harvest efficiency was published in Agriculture journal [8]. Work on detection of anomalies from C-ITS messages using unsupervised ensemble and sliding window concept was submitted to IEEE Transactions on Intelligent Transportation Systems [9].

Also considered was the analysis of speed signatures generated from C-ITS messages with the aim of understanding driving behaviour evolution under a naturalistic driving environment. This work, developed in chapter 5, was presented in IEEE International Conference on Communications (ICC 2021) [10].

The last problem of trajectory linking is investigated by chaining trajectories from C-ITS messages to potential vehicles by considering similarity in movement patterns. This topic is addressed in chapter 6 and a paper was presented in the 2020 IEEE Global Communications Conference (GLOBECOM 2020) [11].

In summary there are a total of five publications and one submitted paper:

1. Juliet Chebet Moso, Ramzi Boutahala, Brice Leblanc, Hacène Fouchal, Cyril de Runz, Stéphane Cormier, and John Wandeto. Anomaly Detection on Roads Using C-ITS Messages. In *International Workshop on Communication Technologies for Vehicles*, pages 25–38. Springer, November 2020.
2. Juliet Chebet Moso, Stéphane Cormier, Cyril de Runz, Hacène Fouchal, and John Mwangi Wandeto. Anomaly Detection on Data Streams for Smart Agriculture. *Agriculture*, 11(11), 2021. ISSN 2077-0472. Doi: 10.3390/agriculture11111083.
3. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Mwangi Wandeto. Streaming-based Anomaly Detection in ITS Messages. Submitted to *IEEE Transactions on Intelligent Transportation Systems*, 2021.
4. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Mwangi Wandeto. Abnormal behavior detection in farming stream data. In: Boumerdassi S., Ghogho M., Renault É. (eds) *International Conference on Smart and Sustainable Agriculture. SSA 2021., Communications in Computer and Information Science*, vol 1470, pp. 44-56. Springer, Cham. June 2021.
5. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, John M. Wandeto, and Hasnaâ Aniss. Road Speed Signatures from C-ITS messages. In *ICC 2021-IEEE International Conference on Communications*, pages 1–6. IEEE, June 2021.
6. Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Wandeto. Trajectory User Linking in C-ITS Data Analysis. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, December 2020.

1.3 Outline

This thesis is organized as follows:

Chapter 2 presents the state of the art. It starts with an exposition of smart agriculture in section 2.2 followed by a review of Intelligent Transportation Systems in section 2.3. Anomaly detection is presented in section 2.4 where various approaches and applications are discussed. Section 2.5 presents trajectory data mining concepts and applications and is followed by section 2.6 on vehicle speed profile analysis. It ends with section 2.7 with a review of trajectory user linking approaches.

In chapter 3 the definitions and methods are presented. It presents a discussion on the performance metrics and measures which will be used in performance evaluation of various algorithms and methodologies used. A discussion on CAM structure and generation rules is presented together with the simulation tools and procedure used to generate simulated CAM data.

Chapter 4 is dedicated to anomaly detection where we propose an Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP) technique. We define an unsupervised data-driven methodology and apply it in three case studies; detection of crop damage in crop dataset, application to GPS logs of combine harvesters and application to C-ITS messages. The main focus is the identification of anomalies that can be linked to crop state/health during harvest, those that have an impact on harvest efficiency and those impacting road safety and efficiency. We utilize techniques based on hypothesis testing on the occurrence of an anomalous movement pattern during in-field harvesting and on the roads.

Chapter 5 discusses the generation and analysis of speed signatures from C-ITS data through the application of trajectory segmentation and statistical analysis. The approaches proposed in this chapter use a real dataset of CAMs collected under the SCOOP@F project [12] with the goal of understanding driving behaviour evolution in a naturalistic driving environment.

Chapter 6 considers the trajectory-user-linking (TUL) problem and applies it to messages generated by vehicles in C-ITS environment. The TUL problem is solved by chaining anonymous trajectories to potential vehicles by considering similarity in movement patterns.

Chapter 7 will conclude this thesis and propose some future works to the presented work.

Chapter 2

State of the art

2.1 Introduction

The advent of location-aware technology such as mobile communication and sensing devices in the big data age has digitized the geographic position of people and things. The properties of generalized data include huge size, rapidly updating, and high value, has resulted in significant changes in people's lives, company operations, and scientific study. This chapter presents a discussion on smart agriculture, ITS, anomaly detection techniques and their application in data streams, agricultural data and road traffic management. It also discusses trajectory mining, speed profile analysis and trajectory linking approaches.

2.2 Smart Agriculture

Agricultural crop production is dependent on the management of elements such as soil, water, pests, as well as climate conditions and unforeseen hazards, for sustainable agricultural output [13]. Pesticide usage in agriculture is a major issue on a worldwide scale, not just in terms of the environment, but also in terms of the health and safety of farmers [14]. Precision agriculture is a strategic approach that collects, processes, and evaluates temporal, spatial, and individual data, as well as other information, in order to support managerial decisions based on estimated variability for increased resource utilization efficiency, productivity, quality, profitability, and stability of agricultural production.

Data-Driven Agriculture, also known as Agriculture 4.0, Digital Farming, or Smart Farming, was created when telematics and data management were integrated with the previously established notion of Precision Agriculture, boosting

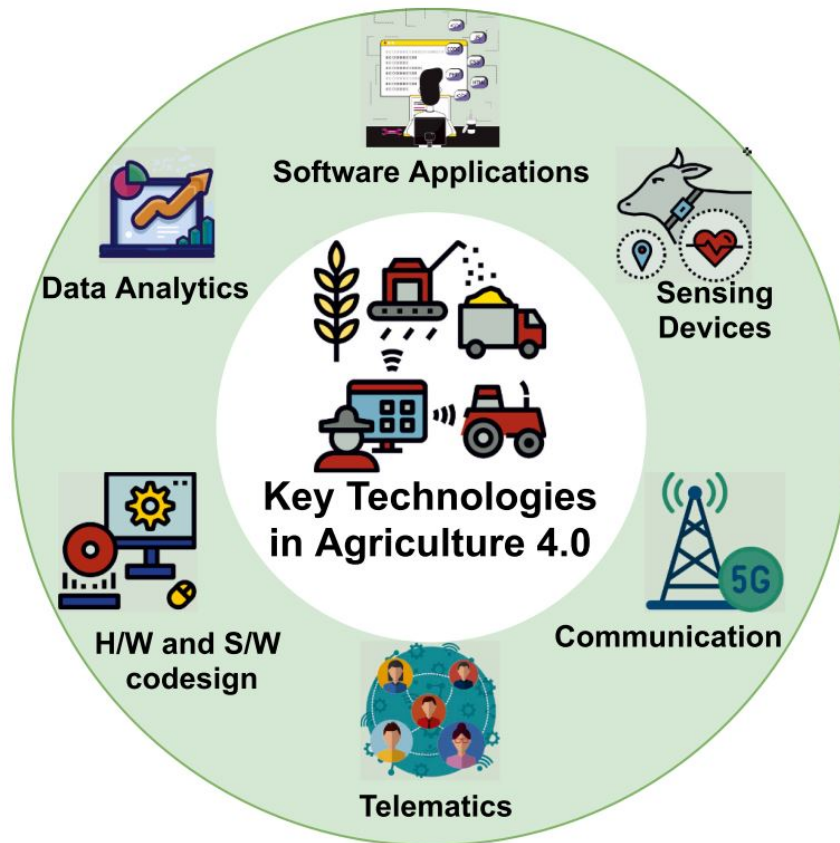


FIGURE 2.1: Representation of key technologies used in Smart Agriculture [44]

operational precision [54]. Smart Agriculture refers to the application of digital methods to innovate, control, and optimize agricultural production systems (as shown in figure 2.1). Human intervention in agriculture is boosted by digital transformation, which aids in reducing effort, implementing particular measures, calibrating the use of chemical products on soil and crops, as well as ensuring and boosting yield. It also aids in the management of all procedures that permit or support agricultural output, such as economic and administrative ones.

Smart Agriculture's goal is to provide solutions that can be used by all farmers, independent of farm size, location, or industry, while leveraging scale effects and keeping costs low. The benefits envisioned from the introduction and integration of technology processes in agriculture are now attributed to increased production and quality efficiency, cost reduction, input optimization, and environmental impact minimization. IoT has the potential to produce solutions for a variety of traditional agricultural concerns, including drought response, yield enhancement, irrigation, and pesticide management. A summary of the applications, services and sensors applied in smart agriculture are shown in figure 2.2.

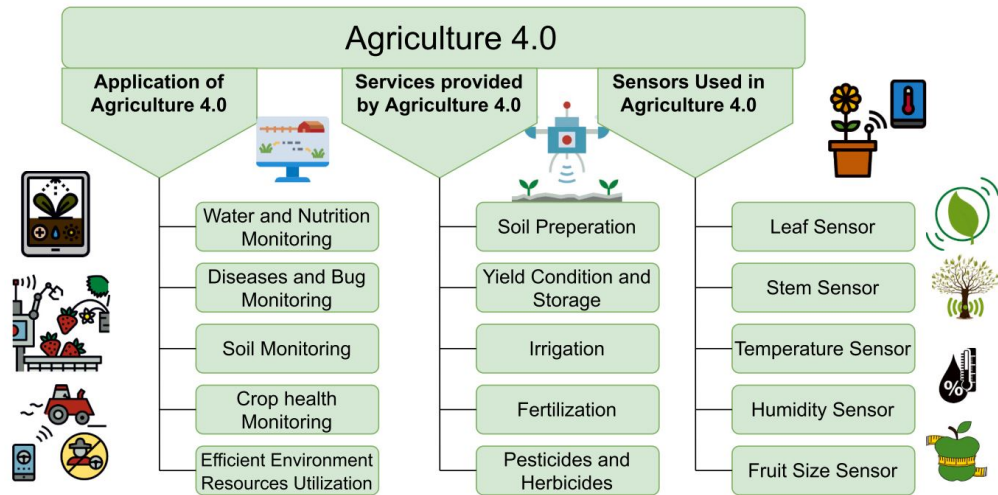


FIGURE 2.2: Representation of applications, services, and sensors used in Smart Agriculture [44]

2.3 Intelligent Transport Systems

Intelligent Transport Systems (ITS) offer substantial advances in transportation system safety, mobility, productivity and environmental conservation [55]. ITS makes use of sensing, analysis, control, and communication technologies. This section highlights the working of Specialized Micro Transportation Systems and C-ITS which are subsets of ITS.

2.3.1 Specialized Micro Transportation Systems

Specialized Micro Transportation Systems (SMTS) are a subset of ITS which comprise of a limited number of vehicles cooperating on a specialized job [15]. SMTS vehicles frequently follow predictable and recognized patterns of activity, which vary depending on the operation at hand. Due to the highly developed and mechanized state of agriculture, it is quite usual for farmers to use numerous vehicles in various essential agricultural operations in order to increase efficiency.

Harvesting is perhaps one of the more engaging of these tasks, since it often involves three types of vehicles: combine harvesters, grain carts, and semi trucks. During grain harvesting a single combine harvester is sufficient for the majority of harvesting activities. However, due to its limited storage capacity and transport speed, a semi truck is generally needed to convey harvested grain from fields to a grain elevator for storage or sale. Since trucks are not adapted to the uneven field surfaces, grain carts are engaged to deliver the grain from the harvesters to the trucks at the edges of the fields. With good coordination between these three

types of vehicles, a reasonable level of efficiency may be achieved in completing the work. ITS can be used to enhance SMTS management, monitoring, and efficiency.

In the developing countries most farmers cannot afford farm machinery, with 55% of farmers in Africa having never used a tractor¹. When tractor services are required, farmers engage dealers who offer the services to them at a fee. Hello Tractor has created a very low-cost tractor monitoring device with computationally intensive software and tools that can be fitted on any tractor. The smartphone application offer visibility into the whereabouts of a tractor. Vodafone's network serves as the platform for IoT connection. This gadget monitors the tractor's condition and communicates any problems that arise [56]. The technology is also being used by tractor dealers to track engine hours and planned maintenance.

In an effort to mitigate fatalities caused by farm vehicles, standards group ETSI presented the first tractor connected to a car using IoT technology in France². To avoid collisions, the tractor sends a warning message to vehicles at a distance of 1 km using an ETSI-standardized IoT communication protocol. The car follows the ITS G5 requirements for transmission of CAM and DENM messages and the 802.11p V2V standard, whilst the tractor complies to the oneM2M (One Machine-To-Machine Partnership Project) protocols designed for IoT.

2.3.2 Cooperative Intelligent Transport Systems

The development of Cooperative Intelligent Transport Systems (C-ITS) focuses on improving safety, comfort, traffic and energy efficiency. The idea behind C-ITS is connectivity of vehicles such that drivers have a good knowledge of the prevailing traffic conditions on the road [16]. C-ITS protocols are standardized by European Telecommunications Standards Institute (ETSI)³ in Europe, and the European Union is pushing for the development of C-ITS in Europe. These systems are deployed in collaboration with road operators, automobile manufacturers, telecom operators, academia, and other industrial providers. The goal of this standardization is to lay the groundwork for vehicle communication interoperability.

C-ITS communication is accomplished through two network types: WiFi network utilizing the protocol 802.11p / G5 [17] which is primarily based on the

¹Africa's IoT Uber for tractors: <https://www.iotworldtoday.com/2021/07/06/africas-iot-uber-for-tractors-highlighted-at-evolution-expo/>

²World's first connected tractor:<https://www.eenewseurope.com/news/iot-tech-enables-worlds-first-connected-tractor>

³ETSI: <https://www.etsi.org/>

broadcast message transmission and cellular network. To transmit and forward messages, three different types of stations are used:

- The On-Board Units (OBU) that are placed in the vehicles.
- The Road Side Units (RSU) that are infrastructures located along the roads.
- The Central Station (ITS-C) architecture has a global network view and is utilized to manage cellular communication of events.

C-ITS exploits the use of V2X communication where by, data sharing is done using Vehicle-to-Vehicle communication (V2V) and Vehicle-to-Infrastructure communication (V2I) [57]. The G5 channel is used for communication between OBUs in V2V mode. V2V communication enables vehicles to communicate wirelessly about their speed, position, and heading. V2V communication technology enables vehicles to send and receive omni-directional signals (up to 10 times per second), giving them 360-degree “awareness” of other vehicles. Vehicles with the right software (or safety apps) can use the messages from nearby vehicles to detect possible collision hazards as they arise. To warn drivers, the system can provide visual, tactile, and auditory signals or a mix of these alerts. These warnings provide drivers the chance to take action in order to avoid collisions. These V2V communication signals may identify threats masked by traffic, topography, or weather and have a range of more than 300 meters.

V2V communication complements and improves existing collision avoidance systems that rely on radars and cameras to identify potential collisions. This new technology not only assists drivers in surviving a collision, but it also assists them in avoiding a collision. V2V communication technology might be used in a variety of vehicles, including automobiles, trucks, buses, and motorcycles (as shown in Figure 2.3) .

The cellular network is utilized by vehicles to convey status information, automated events, and relay events created by other vehicles that do not handle cellular connections to the ITS-C during V2I communication. The G5 network serves the same function, but using the RSU instead (as shown in Figure 2.4).

C-ITS provide real time information on individual vehicles through the use of CAMs. While DENM contains information related to an event that has potential impact on road safety or traffic condition (e.g. road works, accidents, vehicle breakdown etc.) resulting in prompt and active incident prevention. However, the impact of C-ITS services on road networks is yet to be fully felt and evaluated



FIGURE 2.3: Representation of V2V communication

due to the novelty of the concept. This can be attributed to the current low penetration rate of C-ITS equipped vehicles and their compliance rate [2].

2.4 Anomaly detection

Anomalies are “patterns in data that do not conform to a well-defined notion of normal behaviour” [18]. Anomalies are also referred to as outliers, novelties, deviations, rarities or interesting events. Anomaly detection is the task of identifying observations that do not follow the expected pattern. In the presence of labeled data, anomaly detection can be done using supervised learning techniques as a binary classification task, with data instances being either normal or abnormal. However, due to the scarcity of labeled data and the rarity of anomalous events, this is rarely the case. Since large amounts of unlabeled data are accessible, most anomaly detection approaches use unsupervised learning techniques [18].

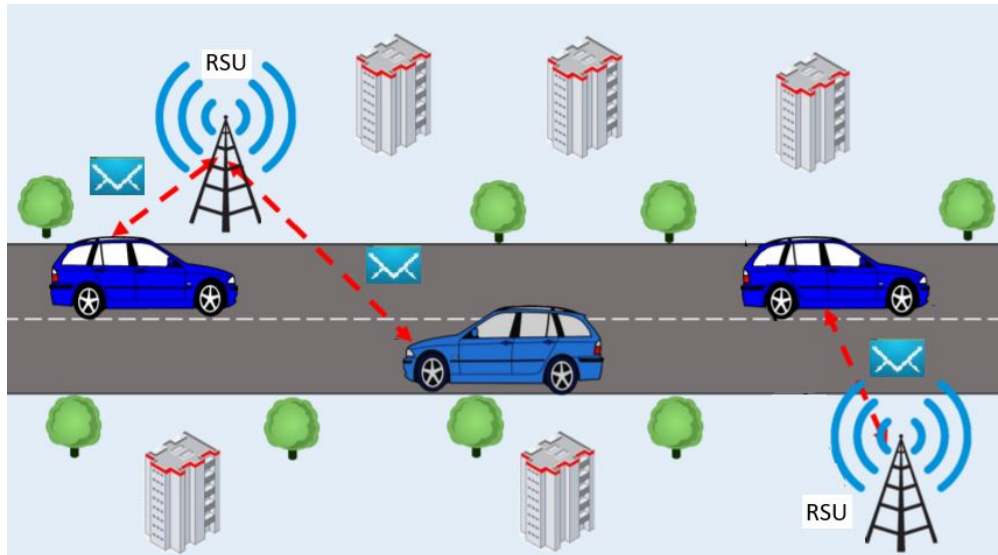


FIGURE 2.4: Representation of V2I communication

Environmental surveillance, fraud detection, structural fault detection, intrusion detection, and disease diagnosis are only a few of the fields where anomaly detection methods have been used. Anomalies can be roughly divided into two categories: (i) incorrect data produced as a result of device failure or system faults, and (ii) irregular data reflecting events that are anomalous but actually occurred [58]. Surprising data is specific to a specific case and as such, using generic methods to identify these anomalies is very difficult.

Techniques for detecting anomalies are typically based on finding occurrences that do not adhere to what is generally considered normal or expected behaviour. According to [59], the challenge in anomaly detection lies in the fact that: (i) the line between normal and abnormal behaviour is blurry; (ii) normal behaviour may evolve into an abnormal depiction in the future; (iii) anomaly detection techniques are not easily transferable across domains due to differences in applications and concepts; (iv) presence of noise in the data makes distinction between noise and true anomalies challenging.

2.4.1 Anomaly types and detection techniques

According to [18] anomalies can be categorized as point anomalies, contextual anomalies and collective anomalies. Individual data elements that are inconsistent or unusual in comparison to all other data elements are called point anomalies. Point anomalies are also known as "outliers" [60]. They are the most fundamental and well-studied types of anomalies. For Example, a vehicle is fitted with sensors which record the vehicle speed as it moves on the road. The captured data has

a regular shape with certain raise/fall in the speed value. At some point, a high raise or low fall in the data can be considered an abnormal behaviour. Figure 2.5 illustrates the speed evolution of the car over time with a sharp increase and fall in speed at the fourth time point representing a point anomaly.

Data items that are regarded uncommon or abnormal in a certain context are known as contextual or conditional anomalies. Anomalies of this nature are widespread in time series data streams. For example, heavy traffic on a roadway during business hours is typical, but at midnight it is contextually abnormal traffic behaviour [45]. Such a circumstance might happen as a result of an accident, disabled vehicles, poor vision owing to foggy weather, or some other reasons. In this form of anomaly, we must also examine other dependant factors to determine if it is an abnormality or regular routine behaviour. When looking at the context, contextual attributes (such as time of day, season, and location) and behaviour attributes describe each data element [18]. Figure 2.6 illustrates distribution of number of cars on the road at different times of the day. Based on the time, the first two peaks are normal while the third peak is abnormal and is considered a contextual anomaly.

A collection or sequence of linked data components that are inconsistent with the entire data set is referred to as a collective or group anomaly. Individual data pieces in collective anomalies may not be abnormal in and of itself, but may constitute a group anomaly when coupled with additional elements. In cardiac behaviour monitoring, a single time interval observation is insufficient to detect cardiac behaviour, but cumulative signals can indicate normal or abnormal behaviour. Figure 2.7 illustrates collective anomaly for heart rate monitoring signal [46].

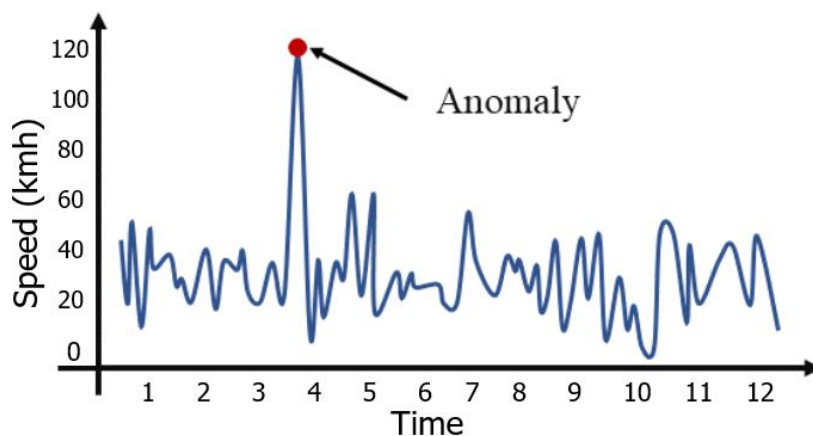


FIGURE 2.5: Illustration of a point anomaly detection

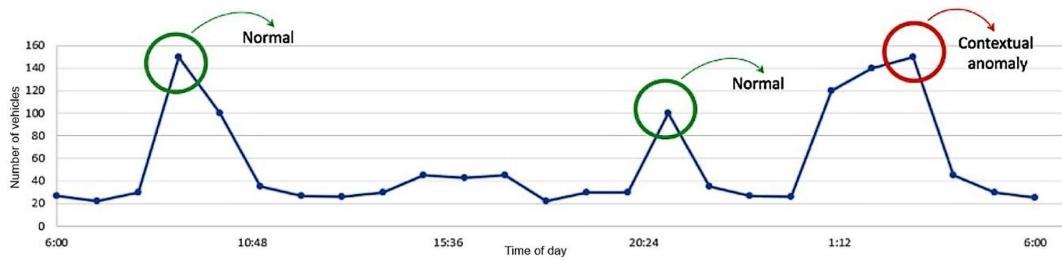


FIGURE 2.6: Illustration of Contextual anomaly detection [45]

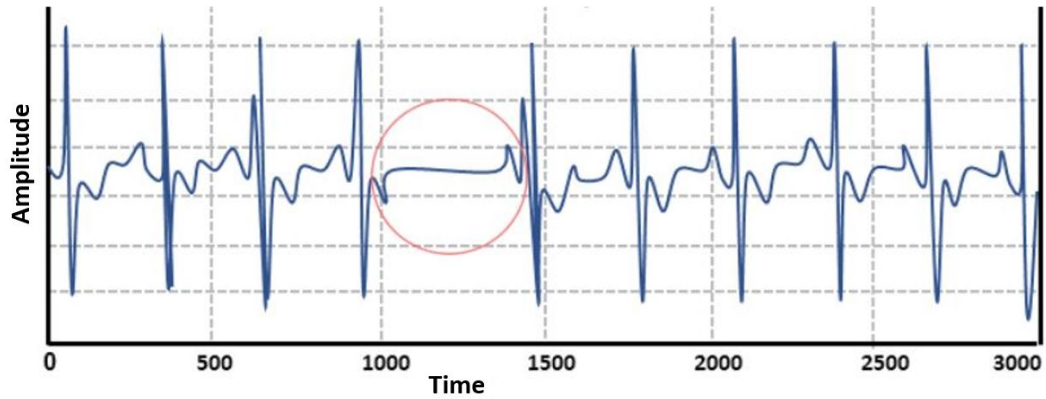


FIGURE 2.7: Illustration of Collective anomaly detection [46]

A great deal of research has gone into the development of a wide range of anomaly detection algorithms which can be broadly categorised as: Classification based, Nearest Neighbour based, Clustering based, Statistical methods, Information Theoretic based, Spectral and Graph based methods [18], [47]. These categories are summarised in figure 2.8 together with the workflow followed in anomaly detection.

2.4.1.1 Statistical approaches

Histogram-based Outlier Score (HBOS) [61] is based on the assumption that features are independent and so computes outlier scores by creating histograms for each feature. HBOS does not require data labeling and provides a quick computation time. Computing time is very important, particularly for C-ITS where there is a very large amount of data to be analyzed to detect anomalies. Anomalies are identified by modeling the detailed characteristics of generated histograms of different traffic features [62] and detecting deviations from the normal road traffic data. When considering numerical elements two types of histograms can be build: static bin-width and dynamic bin-width histograms.

The histograms are normalized in such a way that the maximum height of the bin will be equal to one, this is to ensure an equal weight of each feature. Then,

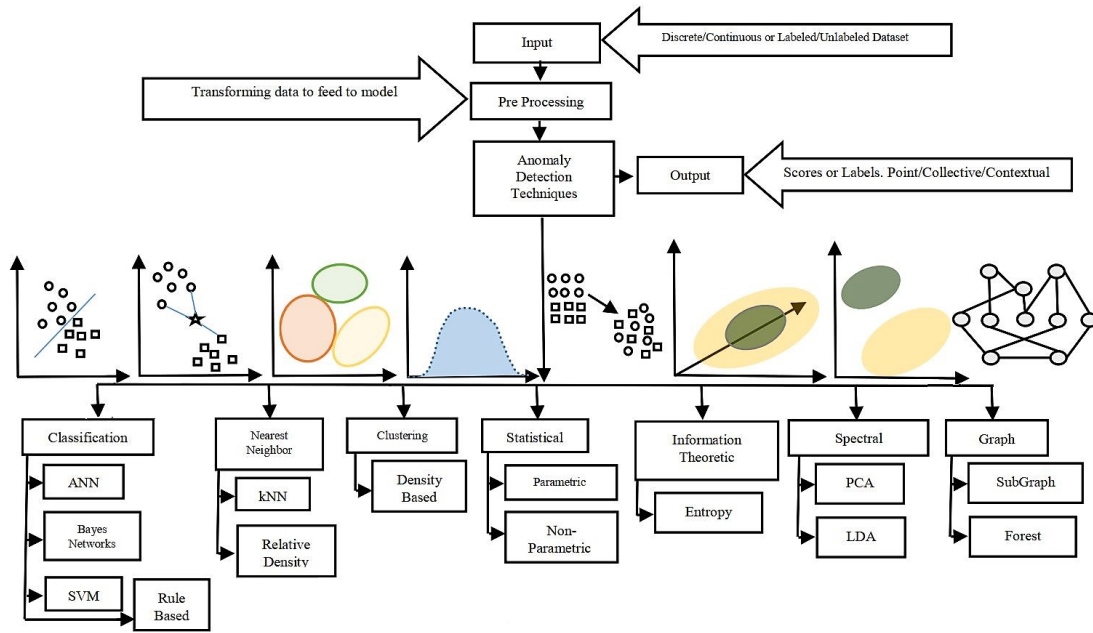


FIGURE 2.8: Anomaly detection workflow and anomaly detection techniques [47]

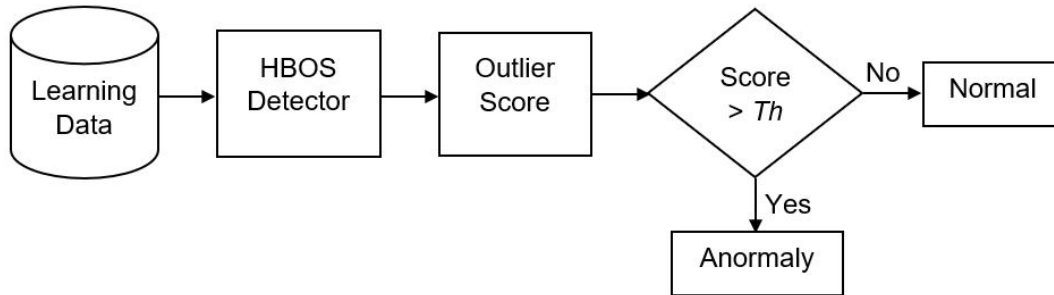


FIGURE 2.9: HBOS workflow for anomaly detection

measured values are reversed such that there is a high score for anomalies and a low score for normal instances. This is done to reduce the sensitivity to errors due to floating point precision which results in extremely unbalanced distributions causing very high scores. The HBOS of every instance x is calculated using the corresponding height of the bins where the instance is located:

$$HBOS(x) = \sum_{i=0}^d \log\left(\frac{1}{hist_i(x)}\right) \quad (2.1)$$

where d is the number of features, x is the vector of features, and $hist_i(x)$ is the density estimation of each feature instance. The workflow of HBOS is shown in figure 2.9.

The scoring part of HBOS generates values which describe the degree of “out-lierness” of each data instance within the learning dataset. The final step is the thresholding step which gives the decision label on whether an element is a normal instance or an anomaly based on a threshold parameter Th . Parameter Th can be set using statistical deviation measures like standard deviation, Median Absolute Deviation (MAD), quantiles, running a moving average on the data within a specific window during stream analysis etc. An instance for example, can be considered an anomaly if it’s score is greater than three times the standard deviation. Another technique is to perform a ranking on the scores such that a top_k algorithm returns the k most abnormal observations.

Another method for detecting anomalies is to assume that data is created by a certain probability distribution and classify points with low probability density as anomalous. This may be accomplished in generalized multivariate analysis (elliptical distribution) by determining the Mahalanobis distance from each point to the mean and classifying anomalies as points with distances greater than a certain threshold. Minimum Covariance Determinant (MCD) [63] is a very reliable estimate for multivariate anomaly identification due to its resistance to outlying observations.

Given a multivariate dataset represented by an $n \times p$ matrix such that n is the number of instances and p is the number of features. The first step in the derivation of the MCD estimator is to compute the determinant of the covariance matrix. A subsample of h observations (with $n/2 \leq h \leq n$) is obtained using a sample of n data points, such that the generalized variance computed on h is minimized. The following location and scatter estimates are defined by the MCD estimator [64]:

1. $\hat{\mu}_0$, the mean of the h observations with the smallest feasible determinant of the sample covariance matrix.
2. $\hat{\Sigma}_0$ is the associated covariance matrix multiplied by a consistency factor c_0 .

The mean and the covariance matrix are applied in the computation of the robust distance for a point x defined as [64]:

$$RD(x) = d(x, \hat{\mu}_{MCD}, \hat{\Sigma}_{MCD}) \quad (2.2)$$

where $\hat{\mu}_{MCD}$ is the MCD estimate of location and $\hat{\Sigma}_{MCD}$ is the MCD covariance estimate.

The reasoning behind minimizing the determinant is that the determinant of a covariance matrix defines how broad the distribution is. As a result, MCD picks the most closely distributed part of the data. This is done to rule out outliers, which are more likely to be located further away from the remainder of the data and thus reduce the masking effect caused by deviating observations [65].

2.4.1.2 Ensemble-based approaches

Isolation Forest (IForest) [66] anomaly detection algorithm is applied in the detection of outliers in high dimensional data. IForest is an unsupervised algorithm that does not require labelled data and does not presume data distribution. It is also non-parametric and performs well on normal unbiased data with few noise points [60]. This makes it suitable for anomaly detection in C-ITS data since the data is unlabelled with no known distribution a priori. IForest is constructed from a forest of random distinct isolation trees (*itrees*). Due to the fact that anomalies are rare and different from normal instances, a tree can be constructed to isolate every instance. The first step of IForest involves the construction of a forest of random *itrees*. In an *itree* data is recursively partitioned such that in each iteration, an attribute is randomly selected and a split point is randomly chosen between the minimum and maximum values of the attribute. The recursive partitioning isolates the instances into nodes with fewer and fewer instances until the points are isolated into singleton nodes containing one instance.

The second step is the scoring phase where IForest assigns an anomaly score for every observation in the dataset. A decision is made at each tree on whether an observation is normal or not. Due to their vulnerability to isolation, anomalies are isolated closer to the root of the tree, while normal points are located at deeper ends of the tree. Thus, the distance of the leaf to the root is used as the outlier score. The final score is achieved by averaging the path lengths of the data points in the different *itrees* of the isolation forest. Given an instance x , the anomaly score is defined as:

$$c(n) = 2H(n - 1) - (2(n - 1)/n) \quad (2.3)$$

$$E(h(x)) = \sum_{i=1}^t h_i(x) \quad (2.4)$$

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2.5)$$

where $E(h(x))$ is the average path length of sample x over t itrees. $c(n)$ is the average path length of unsuccessful search in the Binary Search Tree and $H(i) = \ln(i) + \gamma$ (γ is Euler's constant). Based on the anomaly score s , the following conclusions can be made [67]:

1. If instances return $s(x,n)$ very close to 1, then they are anomalies,
2. If instances have an $s(x,n)$ less than 0.5, then they are considered normal instances, and
3. If all the instances return an $s(x,n)$ of 0.5, then there is no distinction between normal and anomalous instances.

Robust Random Cut Forest (RRCF) [68] is an adaptation of Isolation Forest for stream, that considers concept drift and tree evolution process to define a score of isolation. The anomaly score is determined by how much a new point alters the tree structure. As a result, RRCF is less sensitive to sample size. A robust random cut data structure is used as a synopsis / sketch of the input stream. RRCF preserves pairwise distances during anomaly detection.

LSCP (Locally Selective Combination in Parallel Outlier Ensembles) [48] is an unsupervised detector which defines a local region around a test instance using the consensus of its nearest neighbours in randomly selected feature subspaces. The implementation of LSCP uses an Average of Maximum strategy where a homogeneous list of base detectors is fitted to the training data, then a pseudo ground truth is generated for each instance by selecting the maximum outlier score. It finds and integrates the finest detectors in the local region. By training base detectors on the entire dataset and emphasizing data locality during detector combination, LSCP examines both global and local data relationships. Its strength is that it can measure the degree of local outliers. The process flow of LSCP is summarized in figure 2.10.

A collection of k one-dimensional histograms is used by the Lightweight online detector of anomalies (LODA) [69] to identify anomalies. The probability density of input data projected onto a single projection vector is approximated by each histogram. Because LODA's output is proportional to the sample's negative log-likelihood, the greater the anomaly value, the less likely the sample is. This approach may be used to rank features according to their contribution to the sample's anomalousness since each histogram with sparse projections in LODA provides an anomaly score on a randomly generated subspace. LODA works effectively with data streams and isn't influenced by missing values.

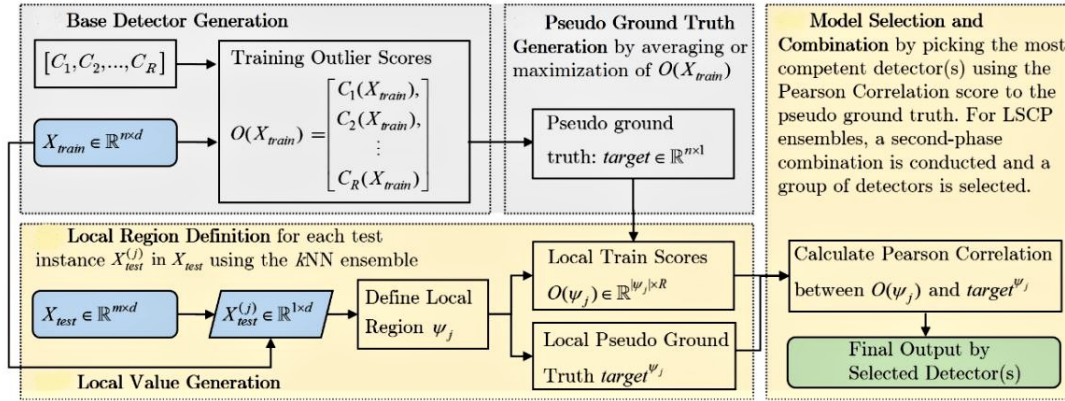


FIGURE 2.10: LSCP flow chart. Steps requiring re-computation are highlighted in yellow; cached steps are in gray [48]

2.4.1.3 Density-based approaches

Clustering and nearest-neighbour algorithms are examples in this category. These approaches determine the distance (similarity measure) between two data points using well-defined distance metrics. Local Outlier Factor (LOF) [70] determines how far a sample's density deviates from its neighbours on a local scale. It is local in the sense that the anomaly score is determined by the object's isolation from the surrounding area. The locality is determined by the distance between the k -nearest neighbours, which is used to estimate the local density. One can discover outliers (samples that have a much lower density than their neighbours) by comparing the local density of a sample to the local densities of its neighbours.

The first step in LOF is to calculate the k -distance between a point p and its k -th neighbour. The distance can be measured in any way, however the Euclidean distance is commonly employed (Equation 2.6).

$$d(p, o) = \sqrt{\sum_{i=1}^n (p_i - o_i)^2} \quad (2.6)$$

Given a dataset D and a positive integer k , the k -Nearest Neighbours of p is any data point q whose distance to p is not greater than k -distance(p) (Equation 2.7):

$$N_{k\text{-distance}(p)}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \leq k\text{-distance}(p)\} \quad (2.7)$$

The reachability distance of data point p with respect to data point o is defined using Equation 2.8

$$\text{reach-dist}_k(p, o) = \max \{k\text{-distance}(o), d(p, o)\} \quad (2.8)$$

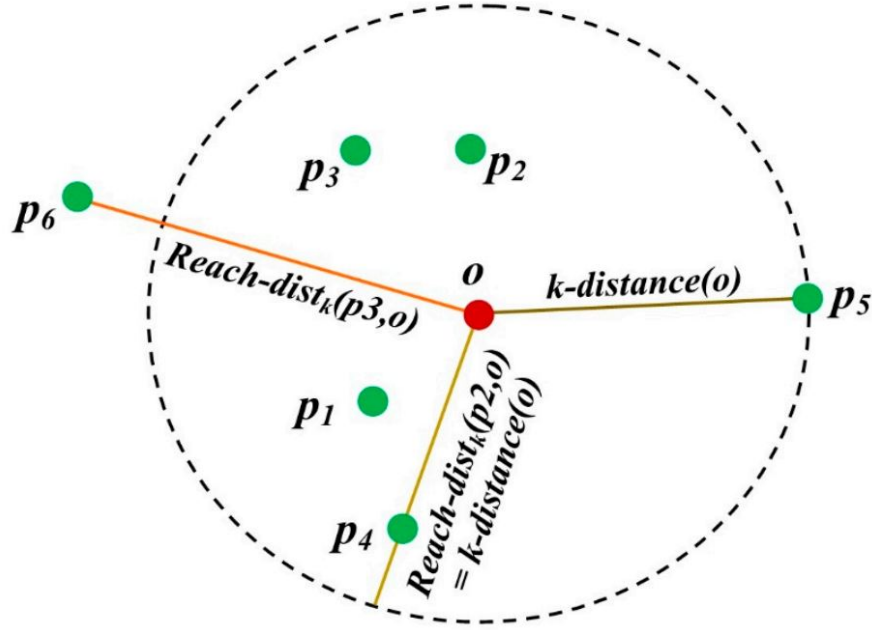


FIGURE 2.11: The reachability distance for different data points p with regard to o , when k equals 5 [49]

Figure 2.11 illustrates the idea of reachability distance with $k = 5$. If object p is far away from object o (e.g., p_6 in the illustration), then the reachability distance between the two is just their actual distance. The real distance is substituted by the k -distance of o if they are "sufficiently" near (e.g., p_1 , p_2 , p_3 and p_4). The rationale for this is that by doing so, the statistical volatility of $d(p, o)$ for all p 's near o may be greatly decreased. The smoothing effect's strength can be adjusted using the k parameter. The greater the value of k , the closer the reachability distances for objects in the same neighbourhood.

The next step is the estimation of the local reachability density (lrd), which is inversely proportional to the average reachability distance of p to its nearest k neighbours (Equation 2.9).

$$lrd_{MinPts}(p) = 1 / \left[\frac{\sum_{o \in N_{MinPts}(p)} reach-dist_{MinPts}(p, o)}{|N_{MinPts}(p)|} \right] \quad (2.9)$$

The LOF is then computed, which is the average ratio of the lrd of point p to the lrd s of its neighbouring points (Equation 2.10). A point is termed anomalous if it is much further apart from its neighbours than they are from each other.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|} \quad (2.10)$$

The main advantage of LOF is that it can successfully identify local outliers.

If a point is close to an extremely dense cluster, it is labeled an outlier. However, because LOF is a ratio, it is difficult to interpret. There is no particular threshold at which a point is considered an outlier. The detection of an outlier is dependent on the problem and the user.

Anomaly detection strategies based on clustering presume that abnormal instances are either in sparse and tiny clusters, far from their cluster centroid, or are not assigned to any cluster at all. Clustering-based Local Outlier Factor (CBLOF) [71] determines the density region of a dataset using a cluster. First, the dataset is clustered using k-means. Then a heuristic technique is utilized to categorize the resulting clusters as large or small. Finally, the outlier score is computed by multiplying the distance of each data point from the central data point in its cluster by the number of data points in the cluster. The distance to the next large cluster is utilized in the case of a small cluster. CBLOF has the problem of being sensitive to the k value in k-means clustering. It also has a large number of parameters and is inefficient at detecting local outliers.

2.4.1.4 Copula-based approach

A copula is defined as “a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform on the interval $[0, 1]$ ”⁴. Copula-based approaches are commonly applied in joint probability distribution modeling due to their advantages in describing heavy tails [72]. Copulas are highly effective in modeling anomalies since they may be thought of as occurring at the tail of the probability distribution. Copula-Based Outlier Detection (COPOD) [73] is a deterministic, non-parametric, fast, and computationally efficient anomaly detection algorithm. It is known to scale effectively in high-dimensional datasets. It is also interpretable, with easy display of anomalies. COPOD first builds an empirical copula, which it then utilizes to forecast the tail probability of each given data point in order to evaluate its level of “extremeness”.

It is a three-stage technique that accepts a d -dimensional input dataset and generates an outlier score vector with values ranging from 0 to ∞ . Stage one fits the left tail and right tail empirical cumulative distribution functions to each dimension in a dataset and computes the skewness coefficient. In stage two, the left tail empirical copula, right tail empirical copula, and skewness-corrected empirical copula values for each row are computed. Stage three computes anomaly scores for each row based on the values calculated in stage two. For each row, the sums

⁴[https://en.wikipedia.org/wiki/Copula_\(probability_theory\)](https://en.wikipedia.org/wiki/Copula_(probability_theory))

of the negative log of the left tail empirical copula, the right tail empirical copula, and the skewness-corrected empirical copula are computed. The anomaly score is the greatest value of these sums in the dataset. A higher anomaly score implies that the data instance has a low probability since it is on the tail of the data distribution. The outlier scores evaluate a row's likelihood in comparison to other instances in the dataset.

2.4.1.5 Linear-based approach

One-Class Support Vector Machine (OCSVM) [74] is an unsupervised learning algorithm. OCSVMs were developed for situations where only one class is known and the objective is to detect anything outside of this class (novelty detection) [75]. It detects the density of the majority class and classifies cases at the density function's extremes as outliers. It repeatedly identifies the maximal margin hyperplane that optimally separates the training data from the origin by mapping input data into a high dimensional feature space (through a kernel).

While OCSVM is efficient at constructing decision surfaces from well-behaved feature vectors, it is poor at modeling variance in big, high-dimensional datasets. As the size of the train set expands, typical OCSVMs impose considerable memory needs and are computationally costly at prediction time, requiring space and time that scales with the number of training points. In realistic, real-world deployments, where extensive training sets are often required to construct correct models when fitting complicated decision boundaries, these memory and computing limits might be prohibitive.

2.4.2 Anomaly detection in Data streams

Streaming data is typically continuously arriving data which is potentially infinite in nature and can be viewed as a multivariate time series. This boundless sequence of arriving data creates a scenario where data may evolve with time resulting in a situation where the most recent data is more relevant in modeling behaviour as opposed to older data [19]. Stream data model [76] can be defined as:

$$Z \equiv \{z(1), z(2), \dots, z(t), z(t+1), \dots\} \quad (2.11)$$

where $z(t) \in \mathbb{R}^N$ for $t \geq 1$

In streaming scenarios where fast processing speed and optimized storage are a priority, anomaly detection algorithms should be able to identify anomalies efficiently. SStream Outlier Miner (STORM), a distance based algorithm for outlier detection was proposed in [19]. Based on the sliding window model, two variants of STORM (exact-STORM and approx-STORM) are proposed to handle outlier queries. In a scenario where the entire window can be allocated in memory, then exact-STORM is used to compute the outliers. In a situation where there is limited memory such that the window cannot fit in the memory, approx-STORM is used to approximate the outliers by introducing effective approximations with a statistical guarantee in exact-STORM.

STORM takes into account the temporal properties of a data point in a data stream. Each data point stays in the sliding window for a certain length of time. Distance-Based Outlier Detection for Data Streams (DBOD-DS) based on adaptive probability density function is proposed in [20]. The probability density function is computed by weighting the data points such that the most recent values are assigned the highest weight with the oldest getting the lowest weight. In this approach there is no assumption on the data distribution, the adjustment of the probability density function is done on-the-fly, thus it is able to capture concept drift.

Anomaly detection in time series data considers the detection of peak values or extreme values where an applicable approach is the use of Peak Over Threshold (POT) [77]. POT method works by isolating all data values which are extreme relative to the rest of the data using a threshold and modelling the tail of all these extreme values by fitting a Generalized Pareto Distribution (GPD). The extreme value theory was adopted by [21], [22] where they propose two algorithms; Streaming Peak Over Threshold (SPOT) for stationary streams and Drift SPOT (DSPOT) for streams with concept drift. In case of a mid-term seasonality, the local peaks will not be detected by SPOT. To mitigate this challenge DSPOT [21] is proposed which applies SPOT to relative values instead of absolute values under the drifting scenario. A moving average approach is applied which uses the windowing concept. In the Internet of Vehicles environment, [78] propose SafeDrive, an online, unsupervised and status aware anomaly detection approach which uses a State Graph model extracted from normal instances as a benchmark for identifying abnormal driving behaviour. Online detection is done by splitting a data stream into segments and comparing each segment with the state graph model. A segment that substantially deviates from the State Graph is marked as an anomaly.

2.4.3 Anomaly detection in agricultural data

Detection of anomalous conditions in the farm environment was done in [23] by application of linear regression to sensor data (soil temperature, soil humidity, soil electrical conductivity, light intensity, air temperature and air humidity). Anomalous data may indicate that the agricultural environment is not conducive for crop growth. The slopes of the regression line were used to determine the trends of each sensor data over time. The outlier detection threshold was set using the Interquartile range (IQR) technique which was used to calculate the upper and lower bounds such that any slopes exceeding the set thresholds were classified as anomalies.

The proposed technique can be improved by combining the environmental data with real-time images of the farm environment in the anomaly detection process. Comparing image variations obtained at different time periods or recognizing particular pest and disease features can also help enhance the accuracy of evaluating abnormal agricultural conditions. This study differs from ours in that we are considering anomaly detection at the tail end of production cycle during harvest with our focus being crop state and harvest efficiency of farm machinery. We also apply Geometric Mean and F1 measure as thresholds for optimum performance of our model as apposed to IQR techniques use in this study.

Deep learning techniques have been applied in agriculture with DeepAnomaly being proposed in [24]. DeepAnomaly uses a combination of background subtraction and Deep Learning for detection of obstacles and anomalies in an agricultural field. It takes advantage of the fact that the visual characteristics of an agricultural field are uniform, and obstacles are uncommon. The major goal is to find things that are far away and severely obscured, as well as unknown object categories. People, barrels, wells, and a distant home were among the observed obstacles. At greater distances, DeepAnomaly identifies persons more precisely and in real time. DeepAnomaly applies image data in anomaly detection and is suitable for real-time applications running on an embedded GPU. This varies with our approach in that we analyse numerical time series data with a purpose of developing an approach that can be applied in general purpose computing systems where memory and processing power are constrained.

Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) are applied in [25] in the detection of occurrence of regional frost disasters using tea frost cases. The study forecasts multiple degrees of hazard for tea tree frost catastrophes, so that producers may not only detect frost incidence and non-occurrence

using the model, but also respond to varying levels of hazard. In [26], two machine learning models, Autoregressive Integrated Moving Average model (ARIMA) and Long Short-Term Memory (LSTM), are used to find anomalies in time series data. The temporal linkages between digital farm sensor data are taken into consideration in the development of a temporal anomaly detection technique. With the goal of finding abnormal data readings, LSTM and ARIMA models are evaluated on actual data gathered from deployed agricultural sensors. It was observed that employing LSTM improves anomaly detection prediction while also necessitating additional training time. LSTM and ARIMA approaches consider important training dataset and time for each case of application, therefore they are not always usable in the case of spatial multivariate data stream with limited computational power allowed.

Precision agriculture is now grappling with the automated detection of crop plots with aberrant vegetation growth. Detecting crop patches that have significantly different phenological patterns than the rest of the crop might help farmers and agricultural cooperatives improve their agricultural practices, disease detection, and fertilizer management. The Isolation forest unsupervised outlier detection algorithm is used in [27] to detect the most abnormal wheat and rapeseed crop parcels within a growing season using synthetic aperture radar (SAR) and multi-spectral images acquired using Sentinel-1 and Sentinel-2 satellites. Heterogeneity problems, growth anomalies, database inaccuracies, and non-agronomic outliers were the four major categories of anomalies studied. Experiments revealed that using both Sentinel-1 and Sentinel-2 characteristics in anomaly detection resulted in the discovery of more late growth abnormalities and heterogeneous parcels. Our approach adopts the use of Isolation forest as one of the base detectors in the ensemble technique with a focus of detecting local and contextual anomalies during crop harvest.

Yield maps are crucial in farmers' decision making when it comes to precision farming, with many farmers having multiyear yield maps [79]. The use of spatial-trend and temporal-stability maps can give valuable information on the spatial and temporal variability of a field. In [79], a contour yield map was generated by interpolating the data into a regular grid using a Kriging [80] linear unbiased estimator. Classification management maps were generated which highlighted homogeneous zones that could be targets for investigation, or sections where inputs might need to be increased or reduced on the basis of yield performance. To identify spatial and temporal trends, a quantitative analysis of yield data from four fields over a six-year period was conducted in [81]. Their previous work in [79]

was modified to separate the temporal effects into two parts: interyear offset (total variations in yield from one year to the next) and temporal variance (the degree of change across time). On the basis of obtained results, interyear variability had the greatest influence on overall yield, with spatial variability being significant within each year and cancelling out over time.

Precision farming has benefited from the integration of high-accuracy GPS technology into farm machinery such as combine harvesters, grain carts, tractors and trucks. Machine data generated by modern agricultural machinery provides crucial information that may be used to obtain valuable insights and forecasts about a farm's operating plan, present status, and productivity. Additionally, based on predictive analytics, farm operators will be able to make adaptive logistical decisions. Visualization of GPS data from farm machinery can provide vital information on operational status and efficiency of the machine. As an example, figure 2.12 presents the movement trajectory of a tractor as it worked in a citrus field and also the path followed to and from the field. From the image, it is clear that a row was skipped during the working of the tractor which can be detrimental to plant growth especially during spraying and fertilizer application.

A Kalman filter and the DBSCAN algorithm were used to detect abnormal activity movement of combine harvesters in [28]. The easting and northing coordinates from GPS logs are iteratively applied to a Kalman filter based on a constant velocity dynamical model. The amount of divergence of the filter estimations from the actual data is calculated using the Kalman filter residual. The Kalman filter residual is a measure of the smoothness of the combine harvester motion, with a greater residual value indicating a rapid shift in motion. The DBSCAN algorithm method is then applied to the engine load, vehicle speed, and calculated Kalman filter residual, which creates clusters indicative of the combine harvester's activity. It was feasible to identify clusters for uniform motion on the road, stationary spots, and nonuniform movements based on the findings. Our approach focuses on in-field harvest efficiency with special emphasis on local contextual anomalies which differ from the activity detection in this study.

2.4.4 Road traffic anomaly detection

Road traffic data anomaly detection can aid in discovery of unusual patterns in traffic which can be analysed further to reveal traffic incidents. Most of the current research focus on traffic incidents caused by short term events like protests, accidents, sporting events, celebrations, vehicle breakdowns etc. [82], [83]. Short term



FIGURE 2.12: Movement trajectory of a tractor in a citrus grove [50]

anomalous events generally resolve within a short time frame. There can also be long term anomalies caused by things like establishment of a mall which attracts a huge number of people resulting in increased volumes on the routes leading to the mall [84]. These long term anomalies generally do not resolve automatically unless urban planning measures are instituted to mitigate their impact.

Traffic congestion and road management are two types of traffic anomalies [29]. Short-term traffic congestion anomalies can last anywhere from a few minutes to several hours. They usually cause a reduction in traffic speed or an increase in traffic volume on the affected roadways. Anomalies in road management are long-term and difficult to resolve. Analysis of traffic anomalies can be carried out by considering the local traffic anomalies or group traffic anomalies. For local traffic anomalies, the road network is partitioned into independent segments, then individual anomalies are extracted per segment. Key features used include attributes like vehicle speed which is extracted from the independent segments. It is then applied to detection algorithms (e.g. clustering based, statistical methods, traffic flow models etc.) to extract traffic anomalies. In the group anomaly category, an

anomaly in one road segment propagates to other adjacent roads and is analyzed by considering the causal interactions among road segments.

In [3] incident detection is done using a real dataset of GPS data from vehicle traces. The authors divide the road network into segments based on the road type, date, time and prevailing weather conditions. Each road segment is assigned a normal average speed range e.g. 80-120 km/h for a motorway type on weekday's peak time. The segments whose average traveling speed is considerably lower than the assigned normal speed are considered abnormal and are extracted. The anomalous segments are further divided into smaller segments so as to isolate the potential incident area. A segment with an incident causing a blockage will have the average speed of the segment in front of it being higher with fewer vehicles. They also consider the behaviour of individual vehicles from the anomalous road segments where they identify vehicles with significantly lower speed than the sub-segment current average speed, stopped vehicles and those moving in a different direction to the traffic flow. The challenge with this approach is that the segmentation process is affected by the accuracy of polygonal line coordinates. Also the accuracy range of GPS affects the distinction of incidence from normal traffic congestion.

Long-term traffic anomaly detection (LoTAD) is proposed in [84] where the main aim is to identify long term anomalous traffic regions in urban cities. In this approach, the road network is partitioned into segments based on bus line data and a real bus trajectory dataset is also segmented based on time slots into temporal and spatial segments (TS-segments) which depict city wide traffic situation. Two trajectory features are extracted, average velocity which defines the traffic condition and average stop time which defines the travel demand. A computation of an anomaly index for TS-segments is done to identify anomalous segments in each bus line which are then consolidated to form anomalous regions. The information extracted from the anomalous regions can be used to offer recommendations for future urban traffic planning.

In [85] Filter-Discovery-Match (FDM) technique is proposed which uses speed patterns to identify incident scenes by partitioning a road network into segments. Speed vectors are generated for consecutive sequences of road segments traveled by a vehicle based on average speed. Real incident data is used to identify incident sections of the road from where incident speed vectors are extracted from vehicles which passed through these sections at the time of the incident. Normal speed vectors are extracted from the road segments by averaging the speed of the vehicles which passed the segments at a specific time frame and were not affected

by the traffic incident. Candidate speed patterns are extracted by computing the speed difference between incident speed vectors and normal speed vector for each segment. These candidate patterns are then clustered using K-means algorithm. Extensive experiments with real taxi data and simulated data revealed that FDM yielded a lower mean time-to-detect (MTTD) than the compared existing techniques.

2.5 Trajectory data mining

According to Liu, Wang and Qu cited in [30], trajectory mining can be viewed as a process of analyzing mobility traces with the aim of discovering spatial, spatial-temporal and behavioural patterns through clustering, classification, anomaly detection, and interesting location detection. Trajectory data mining can also be categorized into the following phases [86]: (a) pre-processing (trajectory compression, stay-point detection, trajectory segmentation and map matching), (b) data management (indexing and storing the data for efficient retrieval) and (c) pattern mining (clustering, classifying, and detecting outliers). The key driving force in trajectory data analysis can be “economic (logistical optimization, customer behaviour analysis, targeted advertising), scientific (animal behaviour analysis, healthcare), administrative (urban planning, criminal investigation), or private” [87]. The present challenge is how to exploit these data to extract useful knowledge and information for improvement of mobility levels [31].

In order to gain useful knowledge from trajectories, the raw points need to be enriched with semantic features, which is essentially a challenging task. One technique for solving this problem is to use experts to annotate the semantic features on the raw trajectories or to let the users attach semantic labels to their trajectories. Another approach is to associate points of interest (POIs) with the location information such that the POIs become the semantic labels [86], [88]. Semantically enriching a trajectory with background information makes querying and analysis simpler and enhances pattern identification [89]. This in turn facilitates behaviour analysis of moving entities. Semantic trajectories can be applied in context-aware computing, trip recommender systems and life experience sharing [86].

In extraction of semantic patterns, the purpose of visits to a location and the time when the particular pattern occurred are important aspects to consider. However, it is challenging to identify the reason for visiting a region due to the fact that the region can cover multiple POIs and most of the time the POIs are not captured as attributes of the trajectory. Also, moving objects generate a lot

of redundant highly sampled data over a long period of time, as a result of the low cost of storage and advances in battery technology. The high sampling-rates over a long period of time is an effective method to increase the probability of capturing more patterns during pattern mining. However, making one-by-one comparisons of the un-simplified raw trajectories is practically impossible and computational intensive. To mitigate this challenge, compression and pruning techniques can be employed during trajectory data mining [90].

Trajectory classification is a process of identifying the class of a moving object based on its movement path. The goal can be to identify a type of vessel, the transportation mode, type of animal or a specific user based on their movement patterns [91]. The key input to a trajectory classification task is a sequence of spatio-temporal points. The main classification process follows three stages [86]: (a) Trajectory segmentation, (b) Feature extraction from the segments, and (c) Building of the classification model (e.g. Dynamic Bayesian Network (DBN), Hidden Markov Model (HMM), and Conditional Random Field (CRF) which consider information from local points/segments and the sequential patterns between contiguous points/segments).

Trajectory segmentation splits a sequence of data points into a series of sub-sequences comprising of homogeneous points based on some defined criteria. A criteria can be derived from the attributes of the trajectory. For example, speed can be used to segment a trajectory by setting a threshold for the difference between the minimum and maximum speed within each sub-sequence. We can also set a threshold for the standard deviation of speed within each sub-sequence. Segmentation of trajectories using spatial density and temporal criteria is considered a cluster-based segmentation problem [92]. A common application of cluster-based segmentation is the detection of stop/move patterns [93],[94]. Segmentation methods can be based on time interval, shape of trajectory and semantic meaning of points (stay point based).

When classifying trajectories, clustering can be performed by assigning similar trajectories to groups (clusters) such that the inter-class similarity is low and the intra-class similarity is high. Clustering facilitates the extraction of collective movement characteristics of objects resulting in behaviour prediction which is used for decision support in location recommendation, destination prediction, weather forecast, urban planning and market research [95]. The current focus of trajectory clustering research is finding appropriate features for trajectory representation, similarity measures and development of algorithms for spatial data clustering [96]. The main challenge is how to identify relevant features that distinguish the class

of a single point, trajectory segment or the whole trajectory and how to select the most discriminate features to be used in building the classification model [97]. A common discriminant feature is the distance between two trajectories or sub-trajectory segments which is computed using a distance measure or metric based on the type of application.

Trajectory similarity encompasses the geometric patterns of moving objects as well as the semantic generalizations derived from the raw trajectories. Several works in literature have considered the geometric or sequential features of trajectories when analyzing user similarity. Similarity among trajectories is often measured in terms of the co-location frequency (feature-based representations), which is the number of times two moving objects appear spatially close to one another. Other approaches for measuring similarity include subsequence similarity metrics such as the length of the Longest common subsequence(LCSS) [98], Edit Distance on Real Sequences (EDR) [99], Common Visit Time Interval (CVTI) [100], Maximal Semantic Trajectory Pattern (MSTP) [101], Multidimensional Similarity Measure (MSM) [102], and Stops and Moves Similarity Measure (SMSM) [103].

LCSS reduces the impact of noisy data by defining distance and matching thresholds. Two points match when their distance is less than a given threshold in all dimensions. However, LCSS ignores possible gaps in sequences, which, for certain problems, results in the same similarity value for different pairs of trajectories. EDR uses an edit distance measure to compute similarity between elements where a match considers all dimensions. Penalties are assigned according to the length of the gaps between two matched sub-sequences resulting in more accurate results than LCSS. CVTI integrates the semantic dimension of stops with temporal dimension. It does not allow heterogeneous data such as stops and moves to be modeled and measured together.

MSTP measures the similarity between two semantic trajectory patterns by considering the frequency at which stops are visited. However, it does not handle multiple data dimensions and does not consider moves between stops. In MSM the similarity score is built upon the matching scores of all pairs of elements that have at least one matching dimension. Partial similarity is assigned according to the number of dimensions in which elements match. It allows definition of different weights for every dimension. It however, ignores the order of stops and does not consider moves. It may assign a high similarity score even if two trajectories are only similar for a small portion of their length. SMSM considers both stops and moves within the trajectory and performs partial dimension matching and partial ordering of stops through assignment of weights. However, estimation of weights

may be challenging for users.

Trajectory data have diverse formats which are unique to application requirements; therefore, different mining techniques and similarity measures are applicable based on the scenario being modeled. When looking at the applicability of similarity measures based on trajectory dimensions, LCSS and EDR require all elements to match across all dimensions, while MSM considers matching pairs in a single dimension. In scenarios where the trajectory data contains outliers LCSS, EDR, MSM and SMSM can be applied since they are robust to noise. When dealing with semantic trajectories MSM and SMSM are good options though, LCSS and EDR can be extended for semantic trajectory mining. When considering applications that use GPS trajectories annotated with stops only or trajectories extracted from social media, the best measure is MSM since it handles sparse data. MSM is particularly useful when one wants to find users who visited the same place at similar times without considering the order of visits. When order of visits is important, SMSM is the most appropriate since it considers the order of the stops. SMSM is also applicable in situations where one wants to extract the most similar paths or most popular routes between stops.

2.6 Speed profile analysis

Vehicle speed and other speed based indicators are commonly used parameters in traffic research for generation of driving profiles. Continuous speed data can be analyzed to reveal speed variation patterns over a particular length of road, acceleration behaviour and generation of total travel time. Based on speed data collected over a relative length of time and distance, information on individual driver behaviour can be extract from analysis of data for a particular vehicle. The research interest may also be on road infrastructure, where the aim is to extract behaviour profiles of road users at intersections, roundabouts, speed humps etc. In this kind of scenario, collective analysis of data from all the vehicles passing through these infrastructure sections will provide a better understanding of movement behaviour.

The primary purpose of analyzing speed profiles is to obtain a better understanding of why drivers react in certain ways to different road/traffic conditions, as well as to identify factors that influence their behaviour. A key feature is the speed of the vehicle, which has a direct consequence on safety, productivity and the magnitude of environmental impact on the traffic ecosystem. Although numerous factors impact collision risk, speed is obviously one of them, with higher

speed being connected with a greater frequency of crashes. When a critical situation occurs, speed affects safety by providing road users less time to react and narrower margins for mistakes [104].

According to [5], speed can be analyzed in three different ways: spot measurements using radar or loop detector, speed profile as a function of time (time-speed profile) or speed profile as a function of space (space-speed profile). The space-speed profiles consider speed against the vehicles position (i.e. speed versus distance traveled from an initial point), and is quite useful in driver behaviour analysis. Space-speed profiles have been applied in the study of driver behaviour in presence of traffic calming measures (speed humps, cushions and chicanes) [105], and at signalized intersections [106].

In [32], outlier detection and machine learning algorithms were used to detect traffic lights, street crossings and roundabouts by analyzing speed and acceleration signals estimated from GPS data. The proposed techniques gave very good results for street crossings and roundabouts. However, for traffic signal detection, lower precision and recall scores were observed across all classification algorithms. A possible extension to this study on infrastructure detection would be to do collective analysis of the data from all vehicles which traveled along a particular road section.

In [107] vehicle speed profiles are used to detect traffic lights using classification techniques. Three methods were used: use of raw speed measurements; use of image recognition techniques; and use of functional data analysis. A comparative analysis of the derived feature approaches shows that functional description of speed profiles with wavelet transformation outperforms the other two approaches. They also show that Random Forests yield an accurate detection of traffic signals, irrespective of the selected feature extraction technique, while keeping a very low confusion rate in stop signs detection.

Under normal driving, vehicle speed reduces to zero or close to zero as a vehicle approaches a stop sign and then picks up after that. This behaviour at stop signs is exploited in [33] where they propose a clustering based stop sign detection method using speed profiles. The deceleration followed by acceleration at stop signs is also used in [34] as a characteristic to detect stop signs from analysis of data generated by on-board car sensors and mobile phone inertial sensors. Extraction of road network properties (intersections and traffic rules) from GPS data was used in [35] to produce high-quality routable maps. Further, it is possible to detect road traffic congestion and incidents in real time using vehicle speed data [36].

In the study of driver profiles, unsupervised machine learning techniques were

applied to a real dataset in [108]. They extracted driving profiles through clustering and obtained communication latency within the globally acceptable range for road safety. Vehicle heading angles were also used in [109] to detect driver profiles within the neighbourhood of Points of Interest (e.g. traffic lights, yield signs, toll zones etc).

In [110], the impact of penetration rate of Dedicated Short Range Communication (DSRC)/ 802.11p and Lane Changing Advisory (LCA) application is studied with a focus of improving travel delay and traffic fluidity. Through simulation, multiple communication densities and vehicle velocities are applied in the evaluation of the impact of different penetration rates on travel speed and traffic fluidity. Their findings show that a 10 percent penetration rate of DSRC is appropriate for LCA to be beneficial for drivers at medium road densities. It is worth noting that V2V communication has an impact on the application efficiency and subsequently on driver behaviour.

2.7 Trajectory-User Linking (TUL)

Trajectory-User Linking is a recent area of research in location based social network applications (LBSNs) [37]. It is motivated by the fact that LBSN applications generate a lot of data which are usually stripped of the user identifiers as a way of anonymizing the data and preserving privacy. On the other hand, linking these trajectories to the users who generated them can provide invaluable information for recommendation systems and identification of criminals through phone signals and check-ins among other applications. Solving TUL is a challenging task due to the large number of user classes and the sparsity of data. A Recurrent Neural Networks (RNN) based semi-supervised learning model, called TULER (TUL via Embedding and RNN) is proposed in [37] which learns the semantic mobility patterns of spatio-temporal data by correlating trajectories to the users who generated them. TULER is designed to identify the dependencies inherent in check-in data and infer hidden patterns of users.

Another semi-supervised learning framework, TULVAE (TUL via Variational AutoEncoder) is proposed in [111] which learns human mobility in a neural generative architecture with stochastic latent variables that span hidden states in RNN. It considers the fact that human trajectories especially in geo-tagged social media are sparse with high-dimensionality and may contain embedded hierarchical semantic structures. TULVAE handles the data sparsity problem by analyzing

large volumes of unlabeled data which is a source of useful knowledge and unique individual mobility patterns.

While considering the heterogeneity of mobility data due to the growing number of location based services and the need for a deep understanding of user behaviour across multiple services, DPLink [112] is proposed. DPLink is an end-to-end deep learning based framework for performing user identity linkage task on heterogeneous mobility data collected from different services with different properties. It is made up of a feature extractor including a location encoder and a trajectory encoder to extract representative features from the trajectory and a comparator to compare and decide whether to link two trajectories as the same user. A multi-modal embedding network and a co-attention mechanism in DPLink handle the low-quality problem of mobility data.

2.8 Discussion

The world's population is continually increasing, necessitating a significant increase in food production. Smart Agriculture tackles agricultural production difficulties in terms of efficiency, environmental impact, food security, and long-term viability. Various approaches and applications have been discussed on how technology is changing the agriculture and transportation landscape.

Anomaly detection was discussed with a highlight of the anomaly types, techniques, applications and challenges encountered in anomaly detection. The presented techniques have drawbacks, for example, some suffer from the curse of dimensionality and are limited to low-dimensional data. Others require selection and tweaking of hyperparameters (for example, the number of clusters for clustering-based models and individual classifier selections for ensemble models). The selection of hyperparameters is sometimes subjective, resulting in varying results.

The anomaly types of interest in this thesis are contextual, and will be investigated further in the agriculture and transportation domains. An unsupervised ensemble technique will be developed to handle multivariate anomaly detection with application of statistical, density-based and tree-based techniques. The idea will be to apply non-parametric techniques so as to minimize the impact of hyperparameters selection.

Trajectory mining was also discussed highlighting the various approaches used to extract knowledge from movement data. Concepts on speed profile analysis were presented highlighting the techniques, features and applications of speed profiles.

A discussion on techniques for linking trajectories to the users who generated them was also presented. Trajectory segmentation and similarity measures are particularly interesting in this thesis and will be applied in driver behaviour analysis. Trajectory segmentation will be applied in generation of speed signatures (in chapter 5) with similarity measures and movement pattern analysis being applied in generation of continuous trajectories through data linking (in chapter 6).

Based on this state of the art, our contributions will propose solutions to anomaly detection, trajectory signature and trajectory linking. Our solutions and their evaluation are based on tools and methods presented in the next chapter.

Chapter 3

Background and methods

3.1 Introduction

This chapter is dedicated to definitions of common terms used in the succeeding three chapters. It also presents a discussion on the performance metrics and measures which will be used in performance evaluation of various algorithms and methodologies used. A discussion on CAM structure and generation rules is presented together with the simulation tools and procedure used to generate simulated CAM data.

3.2 Definitions

Definition 3.1: *GPS log:* Each GPS log l_i is defined by $\langle c_{id}, t, x, y, s, b, a \rangle$ where:

- c_{id} is the combine harvester identifier,
- t is the timestamp of the GPS log,
- x is the longitude of the combine c_k at time t ,
- y is the latitude of c_k at time t ,
- s is the speed of c_k at time t in miles/hour,
- b is the bearing of c_k at time t in degrees
- a is the accuracy of the captured GPS location of c_k at time t

Definition 3.2: *Combine Trajectory:* A raw trajectory consists of time-ordered sequence of n GPS logs of a specific combine harvester such that $T = [p_1, p_2, \dots, p_n]$.

Definition 3.3: *Message:* Each CAM message m_i is defined by $\langle v_{id}, t, x, y, s, h \rangle$ where:

- v_{id} is the vehicle identifier,
- t is the timestamp of the message,
- x is the longitude of the vehicle v_k at time t ,
- y is the latitude of v_k at time t ,
- s is the speed of v_k at time t in metres/second,
- h is the heading of v_k at time t in degrees

Definition 3.4: *Vehicle Trajectory:* A trajectory consists of a sequence of time ordered n messages belonging to a specific vehicle such that $T_k = \langle m_1, m_2, \dots, m_n \rangle$

In this thesis a trajectory is the set of all messages uniquely identified by a single identifier. Another important notion is that of a sub-trajectory. A sub-trajectory is a sequence of consecutive messages that is included in a trajectory.

Definition 3.5: *Sub-Trajectory* T_k^i is a sub-trajectory of T_k if and only if :

- T_k^i is a trajectory: $T_k^i = \langle m_{i_1}, \dots, m_{i_p} \rangle$
- $\forall m_{i_j} \in T_k^i, m_{i_j} \in T_k$
- All consecutive messages in T_k^i are also consecutive in T_k , i.e. there does not exist a message of T_k situated between messages of T_k^i that does not belong to T_k^i :

$$\nexists m_j \in T_k \text{ and } m_j \notin T_k^i \text{ with}$$

$$Rank(T_k, m_{i_1}) \leq Rank(T_k, m_j) \leq Rank(T_k, m_{i_p})$$

3.3 Performance indicators

Anomaly detection research is primarily concerned with the development of new methods and the computational improvement of existing approaches. However, it is critical not to overlook the significance of meaningfully analyzing the performance of existing approaches. Furthermore, it is critical to accurately evaluate

and assess the similarities of different approaches in order to eventually pick the best approach. Anomaly detection is an issue that is intrinsically imbalanced. As a result, the class imbalance problem should be included in the assessment measures of anomaly detection techniques.

The balance between false positives and false negatives in anomaly detection tasks is determined by the application case. False negatives (when no anomaly is discovered while one exists) are usually penalized far more severely than false positives (a false warning). We use Receiver Operating Characteristic (ROC) curves (that is, the true positive rate as a function of the false positive rate) to allow a complete evaluation that is independent of these application-specific trade-offs.

In order to evaluate the performance of the different approaches, we use the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) [113] and the Area Under the Curve of Precision-Recall (AUCPR) [114]. Both indicators are based on the concepts of:

- True Positive (TP): True Positives are the correctly identified anomalies.
- False Positive (FP): False positive are incorrectly identified normal data.
- True Negative (TN): True negative are correctly identified normal data.
- False Negative (FN): False negative are incorrectly rejected anomalies.

The True Positive Rate (TPR), or Recall, is:

$$TPR = \frac{TP}{TP + FN}.$$

The False Positive Rate (FPR) is:

$$FPR = \frac{FP}{FP + TN}.$$

AUC-ROC receiver operating characteristics are TPR and FPR. The higher the AUC-ROC, the better the detection. AUC-ROC is the most popular evaluation measure for unsupervised outlier detection methods [115].

AUCPR uses Precision and Recall. Precision is the fraction of retrieved instances that are relevant [116]. Recall or Sensitivity is the ability of a model to find all the relevant cases within a dataset.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The AUCPR baseline is equivalent to the fraction of positives [117]:

$$AUCPR - baseline = \frac{TP}{TP + FP + FN + TN}$$

The F1 score (F Measure) is the weighted average of Precision and Recall and considers both precision (P) and recall (R) rates. It balances the information provided by precision, which indicates how accurate the model is for correctly categorizing the observations, and recall, which determines how resilient it is, a quality that is obtained when a sufficient number of cases are not missed. The metric is computed as follows:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

AUCPR is a very robust assessment measure that works well for a wide range of classification tasks. It is particularly useful when dealing with imbalanced data, when the minority class is more significant, like in anomaly detection. Therefore, for the purpose of this study, we evaluate the performance of anomaly detection algorithms using these indicators.

3.4 CAM data generation

Cooperative awareness in road traffic implies that road users and roadside infrastructure are aware of each other's position, dynamics, and characteristics. Road users include all types of road vehicles, such as automobiles, trucks, motorbikes, bicycles, and even pedestrians, while roadside infrastructure are road signs, traffic signals, fences and gates. The information to be exchanged for cooperative awareness is contained in the Cooperative Awareness Message (CAM), which is broadcast periodically. A CAM contains information on the originating vehicle's status and attributes. The status information comprises time, position, motion state, activated systems, and so on, while the attribute information includes dimensions, vehicle type, and role in road traffic, among other things.

A CAM is made up of an ITS PDU (Packet Data Unit) header and numerous containers (figure 3.1) that organize the data fields based on the sender's function and the frequency with which they appear in the message. The protocol version, message type, and sender address are carried by the ITS PDU header; the basic container contains the station type and position. The high-frequency container

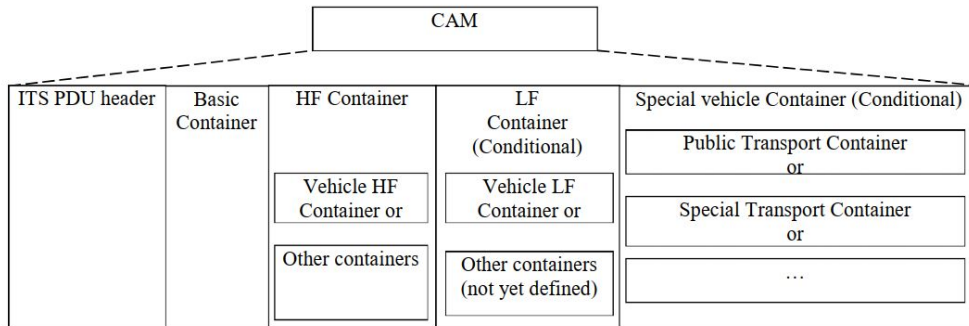


FIGURE 3.1: CAM structure (ETSI EN 302 637-2) [51]

holds mostly highly dynamic data (e.g., vehicle heading, speed, and acceleration) and is transmitted in every CAM. The low-frequency container contains data that has less safety implications (e.g., vehicle role, path history). The special vehicle containers are optionally included if they are required for the sender's function, such as public transportation, dangerous goods, road maintenance, or rescue. The container idea guarantees a flexible message format that can be tailored to the demands of the transmitting and receiving vehicles while reducing the load on the wireless channel [52].

CAM transmission is activated when a vehicle's engine is running. The CAM generation rate as defined in ETSI EN 302 637-2 standard [51] varies between the lower and upper limits of the CAM period $T_{Min} = 100$ ms and $T_{Max} = 1000$ ms respectively (corresponding to a CAM rate of 1 to 10 in 1 second). These limits are controlled by vehicle dynamics, application, and wireless channel congestion state. A CAM is created if the vehicle dynamics exceed the preset criteria for heading, movement, and acceleration as highlighted in condition one of the CAM generation triggering conditions defined in ETSI EN 302 637-2 standard:

1. The time elapsed since the last CAM generation is equal to or greater than T_{DCC} , as applicable, and one of the following ITS Station dynamics related conditions is given:
 - If the absolute difference between the current heading value of the vehicle and the heading value included in the last transmitted CAM by the same vehicle exceeds 4° ;
 - If the distance between the current position of the vehicle and the position included in the last transmitted CAM by the same vehicle exceeds 4 meters;

- If the absolute difference between the current speed of the vehicle and the speed included in the last transmitted CAM by the same vehicle exceeds 0.5 m/s.
2. The time elapsed since the last CAM generation is equal to or greater than T_{APP} and, in the case of ITS-G5, is also equal to or greater than T_{DCC} .

The second condition restricts the minimum and maximum time periods between two CAMs based on the needs of Decentralized Congestion Control (DCC) and the applications to T_{DCC} and T_{APP} , respectively. If the load on the wireless channel is high, the minimum time period is increased, while the application is able to decrease the maximum time period if required based on the safety situation. A low-frequency container and special vehicle container are included in the transmitted CAM if at least 500 ms have passed since the last CAM generation. Figure 3.2 summarizes the rules for CAM generation as specified in the ETSI EN 302 637-2 standard. To guarantee message integrity and authenticity, CAMs include an electronic signature and the relevant certificate. Elliptic Curve Digital Signature Algorithm (ECDSA), which uses elliptic-curve cryptography, is

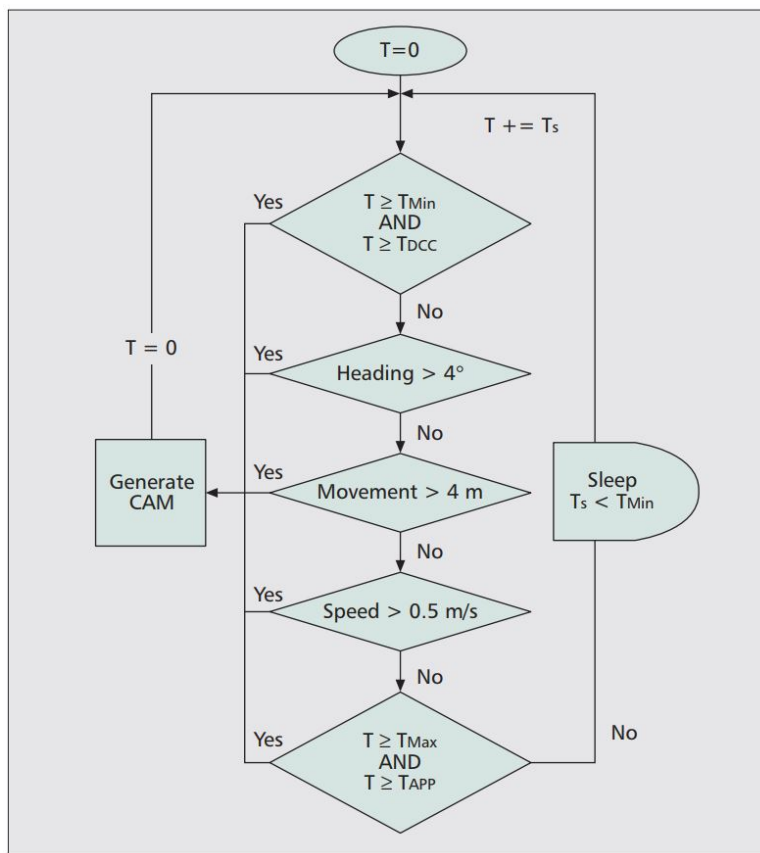


FIGURE 3.2: CAM generation rules (ETSI EN 302 637-2) [52]

selected as the signature algorithm. The recipient of the CAM can then cryptographically validate the message and confirm its temporal validity. The impact of cryptographic data signing is that the data's existence is non-disputable [118]. This means that transmitted CAMs are non-disputable.

3.4.1 Simulation software

The simulation of Vehicular Ad-hoc Network (VANET) applications requires the simulation of both the wireless communication between the vehicles and the mobility of the vehicles. Simulations were carried out using the OMNET++ network simulator [38], and SUMO road traffic simulator [39]. Artery V2X Simulation Framework [40] is used to integrate the network simulator and the road traffic simulator, allowing ease in communication. This integration is particularly important in C-ITS which implements traffic safety and efficiency applications.

OMNET++ is a component-based, extensible, modular C++ simulation library and framework designed particularly for constructing network simulators. It includes an Eclipse-based Integrated Development Environment (IDE), a graphical runtime environment, and a slew of additional tools. Real-time simulation, network emulation, database integration, SystemC integration, and a variety of additional capabilities are supported through extensions. The selection of this software for simulation is due to its functionalities and the widespread use of this platform in the scientific community as well as in industrial settings.

SUMO is an open source software that can be easily integrated with other applications such as OMNeT++ via its Traffic Control Interface (TraCI) (as shown in figure 3.3). This simulator provides for the simulation of intermodal traffic systems such as road vehicles, public transport, and pedestrians. It allows for online interaction with TraCI and the simulation of time-scheduled traffic signals imported or created automatically by SUMO. Another key aspect is that SUMO has no restrictions on the size or quantity of simulated vehicles.

Artery is one of the available frameworks that offers an implementation that incorporates European specification ITS-G5 for vehicular communications. It wraps several entities of an ETSI ITS-G5 stack as OMNeT++ modules, e.g. router, congestion control and security entities. Vanetza provides the implementation of these entities, however, Artery makes them accessible in OMNeT++. Artery enables V2X simulations based on ETSI ITS-G5 protocols like GeoNetworking and Basic Transport Protocol (BTP). Individual vehicles can be configured with a variety of ITS-G5 services using the Artery middleware, which also

offers common facilities for these services. Each vehicle then has its own instance of the ITS-G5 stack (as show in figure 3.4). Some fundamental services are already incorporated, such as cooperative awareness (CAM) and decentralized notification (DENM).

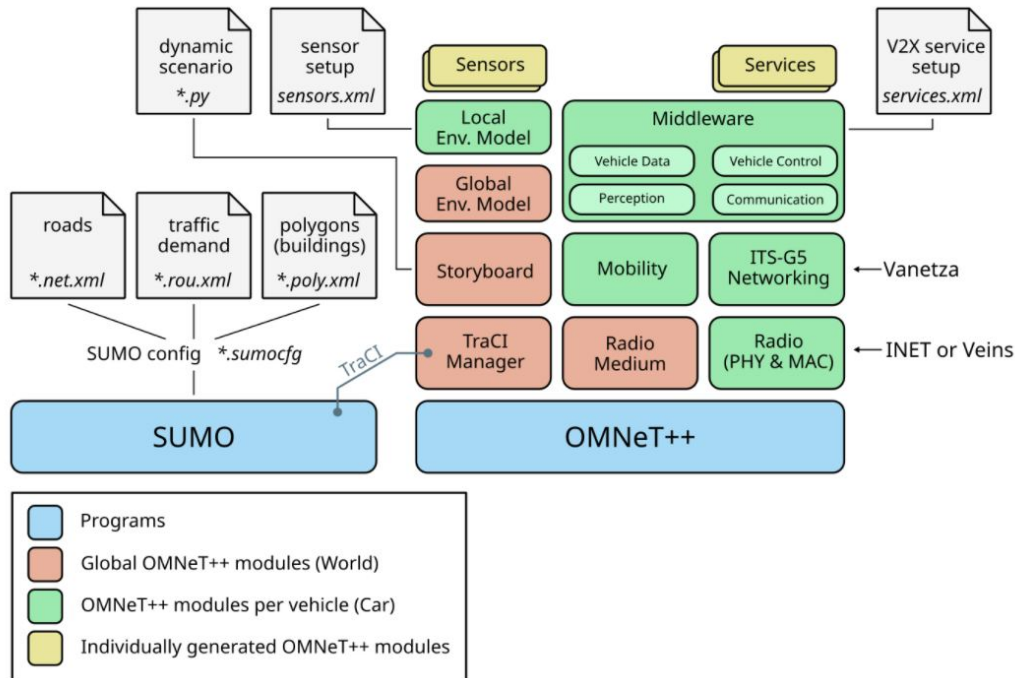


FIGURE 3.3: Major components of Artery simulation framework ¹

3.4.2 Data generation

The simulation featured the mobility of C-ITS equipped vehicles within the city of Reims, France. The simulation parameters were set as described in table 3.1. The first step was the setup of SUMO simulator where the map of the city of Reims in France was extracted from Openstreetmap ² and added to the simulator. Then random trips for 100 vehicles were created. The positions and information of each vehicle were periodically transmitted from the SUMO simulator to the OMNeT++ simulator. The transmitted data was then retrieved from the OMNeT++ simulator by Veins framework and CAM messages created by the Artery framework

The security headers and certificates for the CAM messages were implemented in the Vanetza framework Public Key Infrastructure (PKI) system as defined in the ETSI TS 103 097 technical specifications [119] using the secure message structure.

¹<http://artery.v2x-research.eu/architecture/>

²<https://www.openstreetmap.org/>

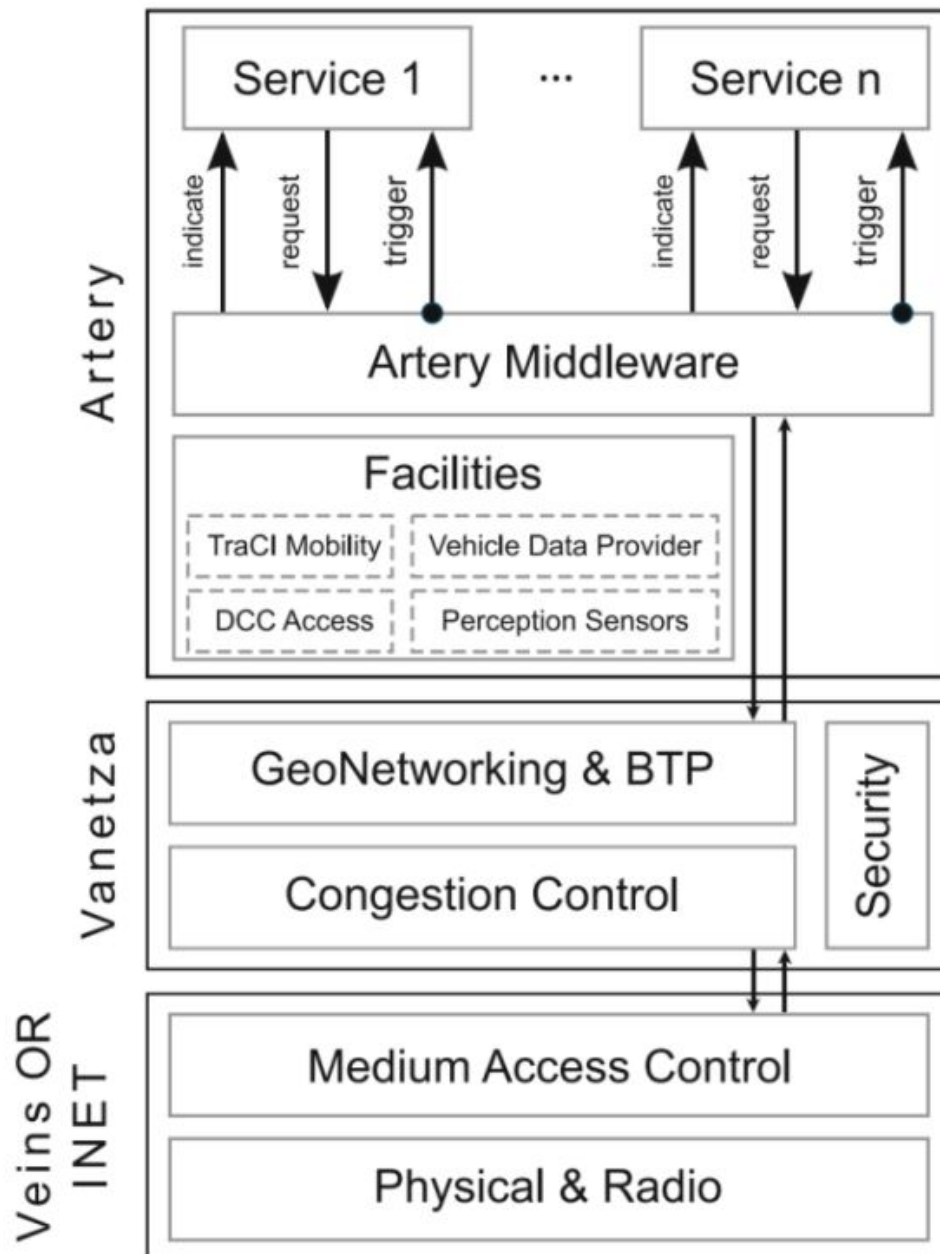


FIGURE 3.4: Artery Architecture [40]

The signature and validation systems of the CAM messages were built using the pseudonyms provided by the PKI system. Each generated CAM was signed by a pseudonym by adding the signature at the end of the message. This signature is used by the validation system at the receiving vehicle to authenticate the received message. The mobility and communications between the vehicles was done in such a way that it mimics the actual movement in C-ITS environment. This was to ensure that the produced results were as realistic as possible. The simulation was run for a maximum of two hours with each vehicle assigned a random route. The CAM messages for each vehicle for the whole simulation were stored in separate

TABLE 3.1: Simulation parameters

Parameter	Value
Network simulator	OMNeT++
Road traffic simulator	SUMO
Framework	Artery (Vanetza, INET, Veins)
Number of nodes	100
Simulation time	7200s
Road map of Reims	10000m * 10000m
CAM message interval	0.1s
Carrier Frequency	5.9 GHz
Number of channels	180
Transmitter power	20 mW

files.

We will now present our contribution to anomaly detection in the next chapter where the simulated data will be applied to our proposed technique and evaluated against other state of the art techniques. The performance measures presented here will be used to evaluate the efficacy of our proposed algorithm.

Chapter 4

Enhanced LSCP Algorithm for anomaly detection on IoT data

4.1 Introduction

The evolution of sensor monitoring technologies and low-cost solutions, together with the incorporation of IoT into everyday life, has resulted in the collection of massive amounts of data [45]. Data streams are vast, continuous, unbounded sequences of data that arrive at a fast rate and have a dynamic distribution (e.g. web searches, sensor data, etc.). Data stream mining is an active research field that has recently evolved in order to uncover knowledge from the vast volumes of continually generated data. Algorithms designed for data streams can handle massive amounts of data. The essential premise of data stream processing is that instances may only be evaluated once since they arrive in a high-speed stream and must then be discarded to create space for subsequent instances. The algorithm analyzing the stream has no control over the sequence of the instances seen and must adjust its model incrementally as each instance is inspected. An additional desirable quality, known as the "*anytime property*" requires that the model be ready to be used at any moment between training instances [120].

Anomaly detection in data streams has three main challenges [121]: (i) because the stream is endless, any off-line learning algorithms that attempt to save the complete stream for analysis will run out of memory; (ii) streams are unbalanced datasets that comprise primarily normal data with a few anomalies thus, detectors requiring labelled data are inappropriate in this situation; (iii) streaming data displays concept drift with time, necessitating the adaptation of anomaly detection models in order to maintain high detection accuracy. Due to their fast speed and

constrained memory, streaming anomaly detection techniques are easy to adapt to real-world applications [122]. Nevertheless, the state of the art stream detection approaches are often oriented to detect a specific type of anomaly.

Outlier detection may be divided into two types: global and local. Global outliers are data points that are beyond the normal range for the whole dataset, whereas local outliers may be within the normal range for the entire dataset but outside the normal range for the data points around them [49]. Anomaly detection is an inherently subjective problem in which the efficacy of detection algorithms varies substantially depending on the problem domain, data characteristics, and types of anomalies [45], [123]. It is also possible that some anomaly detection algorithms will succeed at detecting certain subspaces while others may present poor detection performance [124]. This means that an algorithm can have a domain of expertise in a local domain yet perform poorly over the entire feature space. It is critical to merge the domains of knowledge of each algorithm in order to reduce overall error[125].

The notion of data locality was first presented by the authors of [126], and it was later enhanced in [127] to perform dynamic classifier selection in local spaces of training points. In comparison to static strategies, such as those that simply take a vote based on all base classifier outputs, strategies that dynamically pick and combine base classifiers have produced better results. Ensemble learning approaches merge the predictions from several base models in order to generate more robust, dependable outcomes. To attain unbiased overall detection accuracy with little variance, a desirable anomaly detection ensemble should incorporate the capabilities of various base detectors while carefully combining their outputs to form a robust detector.

Existing anomaly detection ensembles combine several detectors using either parallel or sequential combination structures in an attempt to increase overall detection accuracy by deriving a combined result from the detectors. The purpose of parallel combination structures is to reduce variance, whereas the purpose of serial combination structures is to reduce bias [128]. Nonetheless, incorporating the results of all base detectors may degrade an ensemble's overall performance since, depending on the data sources and underlying rules of a detector, certain detectors may fail to identify anomalies of interest, especially in the unsupervised learning setting.

Unsupervised anomaly detection algorithms aim to automatically identify deviating observations from unlabelled datasets, under some assumptions. Thus, a model performance will be determined by the different characteristics present in

a dataset. This specialization of the models to different characteristics of observations results in varying detection rates. The combination of these individual abilities through an ensemble will be more robust than relying on just one of them [129].

In this study our focus is on contextual anomalies. The concept of a context is inspired by the structure in the dataset, in which each data instance is specified using two sets of attributes [18]:

- *Contextual attributes*: These are used to define the context (or neighbourhood) of an instance. In spatial datasets the longitude and latitude of a location are contextual features. Time is a contextual attribute in time series data that specifies an instance’s placement in relation to the entire sequence.
- *Behavioural/indicator attributes*: These are attributes that relate directly to the process of interest in anomaly detection since they determine the anomalous behaviour. In a spatial dataset describing the average rainfall received for a particular country, the amount of rainfall recorded in any location will constitute a behavioural attribute.

We consider a data driven approach with the objective of detecting anomalies on the fly using unsupervised detection approaches for the detection of local contextual anomalies. We propose an enhancement of an ensemble anomaly detector called Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP). ELSCP is adapted to the streaming context using a pipeline framework which converts data into a stream and passes it to ELSCP using a reference window model which implements a sliding window technique. This adaptation enables the processing of data as a stream which is important in that we can evaluate the performance of our algorithm in the streaming context.

We investigate anomaly detection in three fronts; detection of crop damage during harvest, efficient utilization of combine harvesters and detection of abnormal situations on the road. For each case, we use relevant dataset to identify anomalies and link them to real situations on the ground. We propose techniques based on hypothesis testing on the occurrence of an anomalous movement pattern during in-field harvesting and on the roads. The primary assumption of our analysis is that, “normal instances are far more frequent than anomalies”. The key hypothesis is that, “similar points in a feature space have similar anomaly scores”.

We seek to investigate the following questions:

1. How can the detection rate of anomalies in IoT data streams be improved?
2. Can adaptation of LSCP algorithm yield better results in IoT data stream anomaly detection?
3. How can the adapted technique be applied to real world problems?

We propose the following contributions:

1. We define and investigate the issue of completely unsupervised anomaly ensemble construction;
2. We propose a robust ensemble-based methodology for the detection of anomalies from data streams in smart agriculture and C-ITS context;
3. We apply the proposed technique to crop data with the aim of identifying anomalies that reveal the state of the crop during harvest; and
4. We apply the proposed technique to a data stream of combine-harvester GPS logs with the aim of identifying anomalies that impact harvest efficiency of farm machinery; and
5. We evaluate the proposed technique using a dataset of CAM messages generated in C-ITS environment and compare its performance with state of the art techniques under the streaming context.

4.2 Problem statement

Suppose we have a dataset $Z = \{z_1, z_2, \dots, z_n\}$, where n is the total number of instances. Each instance i of Z consists of both contextual and behavioural attributes. Also, within Z we have a set O of instances which are outliers or anomalies. Our goal is to assign each instance i with an outlierness score S_i such that outliers in O have much higher values than other instances. The outlierness of an instance results from the abnormal behavioural attributes in its context. Given the contextual attributes, there is an underlying pattern that limits the behavioural attributes to some expected values, beyond which an instance will be considered as an outlier.

Definition 4.1: *Contextual outlier:* This is an instance whose behavioural attributes violate the dependent pattern based on its contextual attributes.

Definition 4.2: *Contextual neighbours:* Contextual neighbours of instance i are the instances that are similar to it based on its contextual attributes.

Theoretically, the set of contextual neighbours of instance i is:

$$CN_i = \{j : j \in D \wedge j \neq i \wedge \text{sim}(x_i, x_j) \geq \phi\} \quad (4.1)$$

where

x_i and x_j are contextual attribute vectors, $D = \{1, 2, \dots, N\}$ denotes the set of instances indexes, $\text{sim}(\cdot)$ is a similarity function of two vectors which in our case is a correlation measure, and ϕ is a predefined similarity threshold.

The focus of this study is on unsupervised contextual anomaly detection. It is desirable to create a robust model capable of learning efficiently and accurately predicting an observation as an anomaly when the behavioural attributes are an anomaly given the context. Furthermore, it is preferable for such a model to be sensitive to anomalies in the contextual attributes and to generate meaningful predictions using the best available relevant context. An ensemble based anomaly detection approach with heterogeneous base detectors is applied with the aim of achieving better detection performance in anomaly detection.

4.3 Proposed Enhanced LSCP Algorithm(ELSCP)

The implementations of unsupervised outlier ensembles do not have labels for "outliers" and "inliners". As a result, finding a reliable method for selecting competent base detectors and ensuring model stability is difficult. Traditional unsupervised combination techniques in parallel ensembles are typically generic and global (e.g., averaging, maximization, weighted averaging), but they do not take locality into account. The idea of nearest neighbours in randomly selected feature subspaces is used by LSCP [48] to construct a local region around a datapoint (as discussed in chapter 2 section 2.4.1.2). To create the final model, the top-performing base detectors in this local region are chosen and ensembled. LSCP algorithm was developed to solve two issues: a lack of ground truth and a lack of reliable method for determining suitable base detectors. We developed a new algorithm ELSCP which is an enhancement of LSCP with improvements on how the local region definition is extracted and the selection of competent detectors.

An inherent challenge in anomaly detection algorithms is how to handle variance and bias, especially in ensemble techniques. According to [130], variance is reduced by integrating heterogeneous base detectors by using techniques such as averaging, maximum of average, and average of maximum. A combination of all base detectors, on the other hand, may contain inaccuracies, resulting in greater

bias. As described in Aggarwal’s bias-variance framework, ELSCP combines variance and bias reduction. It improves variance reduction by introducing diversity through the initialisation of heterogeneous base detectors with different hyperparameters. ELSCP also focuses on detector selection based on local competency, which helps identify base detectors with conditionally low model bias.

LSCP uses KD-Tree KNN algorithm with euclidean distance to compute the local region. The efficiency of search trees is determined by the space partitioning approach used, which can be binary or multi-dimensional [131]. KD-tree employs binary splitting, in which only one dimension is considered in each split. For example, in a two-dimensional space, the binary splitting hyperplane is parallel to either the X or Y dimension. Ball tree uses a multidimensional approach in which the split criteria is more flexible and can take into account values from multiple or all dimensions. Furthermore, the splitting criteria constraints define the shape of the resulting partitions, with KD-trees having rectangular partitions and ball-trees having spherical partitions.

When dealing with skewed datasets, binary splits produce long skinny rectangles, which might result in a greater number of backtrack levels during search as well as very unbalanced trees. Furthermore, rectangles and squares are not the optimum forms for partitioning because if a target point falls into a corner of a rectangle, tracing over a large number of nodes around the corner to identify nearest neighbour increases the complexity of the search algorithm [132]. As a result, using metric-trees like the Ball tree, where the space partitioning using hyperspheres is explicitly tuned for the distance function, is the most efficient approach [131], [133].

Our focus in anomaly detection is to detect local contextual anomalies where our contextual attributes are longitude and latitude since we are dealing with spatial-temporal data. The most precise metric for computing spherical distances on the earth’s surface is the harvesine distance. Therefore, to improve the local region definition, we propose to implement local neighbour search using Ball Tree KNN algorithm with Harvesine distance metric.

The computation of Harvesine distance is presented in equation 4.2

$$d = 2r \sin^{-1} \left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)} \right) \quad (4.2)$$

where r is the radius of the earth(6371 km), d is the distance between two points, ϕ_1, ϕ_2 is the latitude of the two points, and λ_1, λ_2 is the longitude of the two points respectively.

Suppose we have a dataset \mathbb{R} which we split into training and test datasets: let $X_{train} \in \mathbb{R}^{m \times a}$ represent the set of training data with m points and a attributes, and $X_{test} \in \mathbb{R}^{y \times a}$ be the set of test data with y points and a attributes. The procedure for ELSCP starts with a heterogeneous list of base detectors D being fitted to the training data. The result of this training is a predicted set of outlier scores O_{train} which is presented in equation 4.3

$$O(X_{train}) = [D_1(X_{train}), D_2(X_{train}), \dots, D_k(X_{train})] \in \mathbb{R}^{m \times k} \quad (4.3)$$

The second step is the generation of training pseudo ground truth (*target*) which is done by selecting the maximum score across all the detectors from O_{train} . The third step is the definition of the local region (ξ) for each test instance, X_{test_i} . It is defined by computing the k nearest neighbours of each instance using Ball tree KNN algorithm with Harvesine distance. This is formalized as:

$$\xi = x_j | x_j \in X_{train}, x_j \in L_{ens} \quad (4.4)$$

where L_{ens} is the collection of a test instance's nearest neighbours subject to Ball tree ensemble criteria.

To define the local region L_{ens} , feature spaces are constructed by randomly selecting t groups of $[d/2, d]$ features. In each group, the k nearest training objects to X_{test_i} are identified using harvesine distance. The training objects which appear more than $t/2$ times are added to L_{ens} . With the local region defined, a local pseudo ground truth ($target_s \in \mathbb{R}^{s \times l}$) is generated by extracting the points in ξ from *target*. The local training outlier score for the test instance is extracted from the pre-computed training score matrix O_{train} using equation 4.5

$$O(X_{train_s}) = [D_1(X_{train_s}), D_2(X_{train_s}), \dots, D_k(X_{train_s})] \in \mathbb{R}^{s \times k} \quad (4.5)$$

The fourth step is the selection of the optimal detector which is done by computing the similarity between the base detector score and the pseudo target using correlation measures. Similarity computation is motivated by the absence of direct and reliable access to binary labels in unsupervised outlier detection. Although it is possible to convert pseudo outlier scores to binary labels, determining an exact conversion threshold is difficult [134]. Furthermore, because unbalanced datasets are typical in outlier identification tasks, using similarity metrics is more reliable [48]. We propose two implementations of ELSCP, one which uses Pearson correlation and weights to compute the final score and the second one uses Kendall rank

correlation scoring technique.

Definition 4.3: *Kendall Rank Correlation:* Let $(x_1, y_1), \dots, (x_n, y_n)$ be a set of observations of the joint random variables X and Y , such that all the values of x_i and y_i are unique [135]. Kendall rank correlation is a statistic of dependence between two variables [136]. It is a measure of the similarity of the orderings of the data when ranked by each of the quantities. Kendall correlation is non parametric and does not make any assumptions on the distribution of the data. It is preferred when the relationship between two variables is none linear and when the normality assumption for the two variables is invalid. Its definition is presented in equation 4.6

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j) \quad (4.6)$$

Definition 4.4: *Pearson Correlation:* Let x and y be two vectors of length n , where \bar{x} and \bar{y} are the means of the vectors. Pearson correlation coefficient is defined as the ratio of the co-variance of the two vectors to the product of their respective standard deviations (presented in equation 4.7). It measures the linear association between two numerical variables. Also, it is applicable for features that are normally distributed.

$$r_{xy} = \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.7)$$

where

n is the sample size; $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ denotes the mean of x ; $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ denotes the mean of y

ELSCP implementation with Kendall rank correlation computes the correlation between the local pseudo ground truth ($target_s$) and the local detector scores $D_i(X_{train_s})$ as $\tau(target_s, D_i(X_{train_s}))$ using equation 4.6. This computation iterates over all the k base detectors. A histogram is built with b equal intervals out of Kendall rank correlation scores, and detectors in the largest bin are selected as competent base detectors for the given test instance. Finally, the selected detectors scores are combined using the average of maximum strategy as $avg(D_t^*(X_{test_i}))$, yielding the final detection score.

ELSCP implementation with Pearson correlation computes the correlation between the local pseudo ground truth ($target_s$) and the local detector scores

$D_i(X_{train_s})$ as $r(target_s, D_i(X_{train_s}))$ using equation 4.7. This computation iterates over all the k base detectors. A histogram is built with b equal intervals out of Pearson correlation scores, and detectors in the largest bin are selected as competent base detectors for the given test instance. Weights are then computed by ranking the Pearson correlation scores. Finally, the selected detectors scores are combined using the weighted average of maximum strategy as $weighted_avg(D_t^*(X_{test_i}))$, yielding the final detection score. The implementation of ELSCP is summarized in algorithm 1, with figure 4.1 showing the flow chart.

Algorithm 1: Enhanced LSCP

Input : the pool of heterogeneous detectors D , training data X_{train} , test data X_{test} , the local region size k

Output: Outlier scores for each instance in X_{test}

Train all base detectors in D on X_{train}

Generate training outlier scores O_{train} with Eq.4.3

Generate pseudo ground truth: $target := \max(O(X_{train}))$

for each test instance X_{test_i} in X_{test} **do**

 Define local region (ξ) by Ball tree kNN ensemble

 Extract local pseudo ground truth $target_s$ by selecting k neighbours in (ξ) from $target$

for each base detector D_i in D **do**

 Get the outlier scores associated with training data in the local region $D_i(\xi)$

if $ELSCP_K$ **then**

 Evaluate the local competency of D_i by computing the similarity between $target_s$ and $D_i(\xi)$ using Kendall correlation with Eq.4.6

else

 Evaluate the local competency of D_i by computing the similarity between $target_s$ and $D_i(\xi)$ using Pearson correlation with Eq.4.7

 Select a group of t most similar detectors and add to the empty set D_t^*

if $LSCP_P$ **then**

 Compute weights by ranking the Pearson correlation scores

return weighted $avg(D_t^*(X_{test_i}))$

else

return $avg(D_t^*(X_{test_i}))$

return $scores$

4.3.1 ELSCP Algorithm using Kendall rank correlation (ELSCP_K)

ELSCP_K implementation applies three base detectors, HBOS, MCD, and isolation forest (IForest). Histogram-based outlier score (HBOS) [61] is based on the assumption that features are independent, and hence computes outlier scores by creating histograms for each feature (detailed discussion in chapter 2 section 2.4.1.1). Modelling the precise features of produced histograms and identifying deviations are used to identify anomalies. HBOS does not require data labeling and does not require any training or learning phase. With scoring-based detection, it also provides quick computation, which is important in our context since we process data streams. The IForest [66] anomaly-detection algorithm is unsupervised, does not assume the distribution of the data, and does not require labelled data. It also performs well on normal unbiased data with few noise points and is nonparametric [60]. IForest is constructed from a forest of random distinct isolation trees (*itrees*) (detailed discussion in chapter 2 section 2.4.1.2). IForest properties inform the choice for application in this study because anomalies are rare and different from normal instances; therefore, a tree can be constructed to isolate every instance.

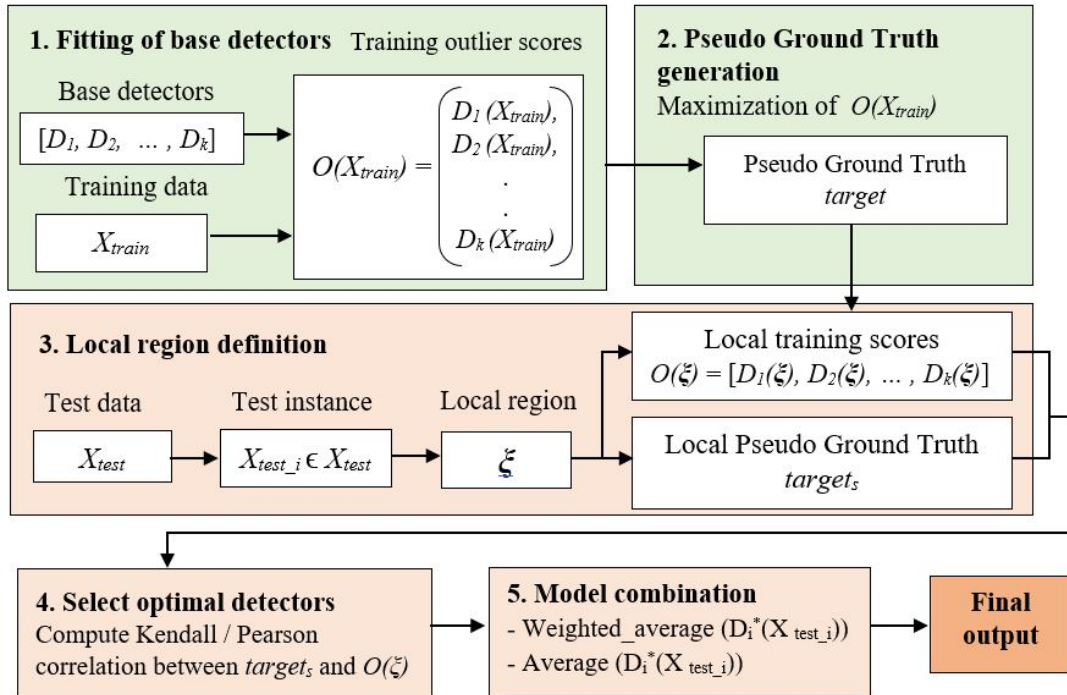


FIGURE 4.1: ELSCP flow chart: The results for steps 1 and 2 are cached; steps 3 - 5 are re-computation for each test instance

Diagnostics are frequently used in combination with a standard fitting technique to detect outliers. Outliers, on the other hand, can impact these methods, causing the fitted model to miss deviating observations (masking effect) [65]. Robust statistics is a useful technique for discovering these deviating observations since it fits the majority of the data to the fitted model and then identifies data points that differ from the fitted model. In this study, we mitigate the masking effect by implementing minimum covariance determinant (MCD) [63] (detailed discussion in chapter 2 section 2.4.1.1), which is a very reliable estimate for multivariate anomaly identification as one of the base detectors in ELSCP_K. The determinant of a matrix and Mahalanobis distances are two multivariate feature statistics on which MCD relies. It searches the dataset for observations with the lowest possible determinant in the classical covariance matrix.

The procedure for ELSCP_K starts with HBOS, MCD, and IForest base detectors being fitted to the training data. A pseudo ground truth for each train instance is generated by taking the maximum outlier score from all the base detectors. We implement Kendall correlation in the model selection and combination phase. For each test instance:

1. Using a ball tree nearest neighbour algorithm with Haversine distance metric, the local region is defined to be the set of nearest training points in randomly sampled feature subspaces that occur more frequently using a defined threshold over multiple iterations.
2. Using the local region, a local pseudo-ground truth is defined, and Kendall correlation is calculated between each base detector's training outlier scores and the pseudo-ground truth.
3. A histogram is built out of Kendall correlation scores, and detectors in the largest bin are selected as competent base detectors for the given test instance.
4. Using the correlation scores, the best detector is selected. The final score for the test instance is computed by using the average of the best detector's local region scores.

4.3.2 ELSCP Algorithm using Pearson correlation (ELSCP_P)

ELSCP_P implementation applies two base detectors, HBOS and LOF. Local Outlier Factor (LOF) [70] determines how far a sample's density deviates from its neighbours on a local scale (detailed discussion in chapter 2 section 2.4.1.3). It is local in the sense that the anomaly score is determined by the object's isolation from the surrounding area. The locality is determined by the distance between the k-nearest neighbours, which is used to estimate the local density. One can discover outliers (samples that have a much lower density than their neighbours) by comparing the local density of a sample to the local densities of its neighbours.

The procedure for ELSCP_P starts with HBOS and LOF base detectors being fitted to the training data. A pseudo ground truth for each train instance is generated by taking the maximum outlier score from all the base detectors. We implement Pearson correlation in the model selection and combination phase. In the computation of the final ensemble score, we implement a weighted average where the weight is computed by ranking the Pearson correlation scores. For each test instance:

1. Using Ball Tree nearest neighbour algorithm with Haversine distance metric, the local region is defined to be the set of nearest training points in randomly sampled feature subspaces which occur more frequently using a defined threshold over multiple iterations.
2. Using the local region, a local pseudo ground truth is defined and the Pearson correlation is calculated between each base detector's training outlier scores and the pseudo ground truth.
3. Weights are computed for each detector by ranking the Pearson correlation scores such that the detector with the best score gets the highest weight.
4. Using the correlation scores, the best detector is selected. The final score for the test instance is computed using a weighted average of the best detector's local region scores.

ELSCP_P is adapted to streaming context by implementing it through a pipeline as shown in figure 4.2. Given a dataset of CAMs the stream simulator converts the data into a data stream. The stream is passed to ELSCP_P through a reference window model which implements the windowing concept. Within the specified window size, a sliding window is implemented such that partial anomaly

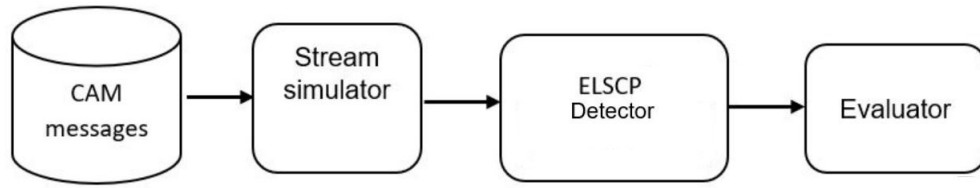


FIGURE 4.2: ELSCP workflow

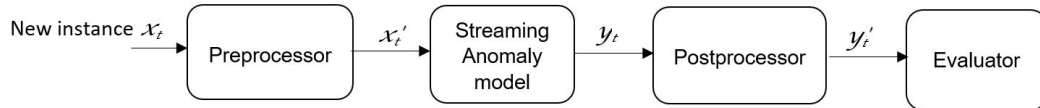


FIGURE 4.3: Anomaly Detection framework

scores are generated for each model. The partial scores are then evaluated and final scores generated for each instance.

A sliding window approach is employed in streaming anomaly detection, in which data samples inside a window are sorted by an outlier score, with highly ranked data samples being labeled anomalies. For ELSCP_P adapted to streaming context a pipeline framework was adopted where each incoming new instance x_t is passed through a pre-processor (unit norm scaler) which transforms x_t into a scaled feature vector without changing its dimensions. The scaled feature vector is then processed by the streaming anomaly detection model which predicts the label y_t for the instance. This predicted label is then passed to the running average post-processor which converts the score to the average of all previous scores in the current window. Figure 4.3 depicts the proposed anomaly detection framework.

The key advantage of using sliding window is that, with the arrival of a new data instance, a sliding window may be modified, resulting in an online and incremental updating process. The update mechanism necessitates the deletion of an old data instance and the storage of a new one, making it computationally efficient. The sliding window contains the latest subset of the dataset at any given time. As a result, a sliding window approach finds outliers based on the most recent subset and addresses the temporal property of data streams.

4.4 Application Scenarios

The implementation of ELSCP is applied to three scenarios. The first scenario is detection of anomalies in crops which can be linked to crops state (i.e. if the crop is alive or is damaged at harvest time). The second scenario is applied to detection

of anomalies which reflect abnormal harvesting behaviour of combine harvesters. The third application is in Cooperative Intelligent Transport Systems where we detect anomalies on the road.

4.4.1 Scenario A: Crop dataset

The first scenario considers anomalies in crop data where the aim is to link the detected anomalies to the state of the crop at harvest, whether it is alive in good health or is damaged. The crop dataset [137] was collected by various farmers at the end of the harvest season spanning a period of three seasons. To simplify the analysis, we assume that all other factors like variations in farming techniques were controlled. This dataset has ten variables as summarized in Table 4.1.

In this study we consider the data for crop type zero(0), which we extract from the crop dataset. The first pre-processing step was to check for missing values in the data which reveals 6397 missing values in Number_Weeks_Used variable. The missing values were filled using Iterative imputation technique [138] which does multivariate imputation by chained equations. In iterative imputation each feature is modeled as a function of the other features. Each feature is imputed one after the other, allowing previously imputed values to be utilized as part of a model to predict future features. The next step was to check for the distribution of the data per attribute, which is summarized in Figure 4.4.

Exploratory data analysis was done on the crop data so as to gain a better understanding of the data through generation of a correlation heatmap (as

TABLE 4.1: Crop Data Description.

Column Name	Description
Id	UniqueID
Estimated_Insects_Count	Estimated insects count per square meter
Crop_Type	Category of Crop(0, 1)
Soil_Type	Category of Soil (0, 1)
Pesticide_Use_Category	Type of pesticides used (1 - Never, 2 - Previously Used, 3 - Currently Using)
Number_Doses_Week	Number of doses per week
Number_Weeks_Used	Number of weeks used
Number_Weeks_Quit	Number of weeks pesticide not used
Season	Season Category (1, 2, 3)
Crop_Damage	Crop Damage Category (0 = alive, 1 = Damage due to other causes, 2 = Damage due to Pesticides)

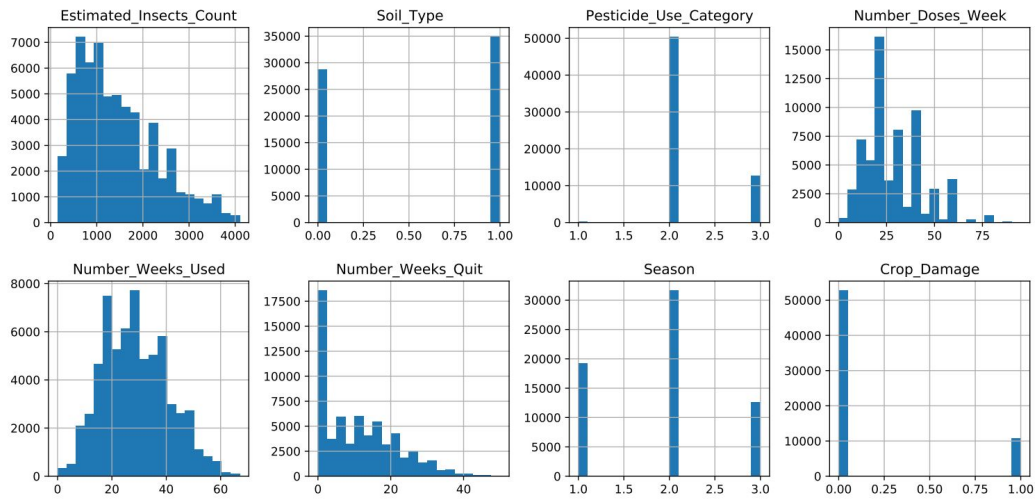


FIGURE 4.4: Histograms showing data distribution per attribute in the crop dataset

shown in Figure 4.5). Based on the heatmap, `Estimated_Insects_count`, `Pesticide_use_category` and `Number_weeks_used` are positively correlated with `Crop damage`. This indicates that the state of the crop at harvest time will be greatly impacted by the presence and number of insects, the type of pesticides used to control these insects and the duration of exposure to the pesticides. On the other hand, `Number_weeks_used` is positively correlated with `Estimated_Insects_count` and `Pesticide_use_category`. High negative correlation exists where `Number_weeks_Quit` is negatively correlated with `Pesticide_use_category` and `Number_weeks_used`. In addition, a box-and-whisker plot (boxplot) was generated to determine the existence of outliers in the dataset. Each graph's box depicts the attribute range between the first Quartile (25th percentile) and the third Quartile (75th percentile), while the lines within each box depict the median value. The whiskers extend to the most extreme data points that are not regarded outliers, whereas outliers are represented separately using the 'o' symbol. Figure 4.6 shows that except for soil type and season, all other attributes have outliers.

4.4.2 Scenario B: Combine harvester GPS logs

We used a GPS dataset [139] collected using Nexus 7 tablets on a farm in Colorado U.S.A during 2019 wheat harvesting season. A GPS recorder application was kept running for each involved vehicle (combine harvesters, grain carts and trucks) during harvesting. The combine harvester moves in a straight line at a near constant speed in a normal combine harvester operation (whether harvesting or traveling between fields). The sampling frequency of the data was 1 sample (GPS

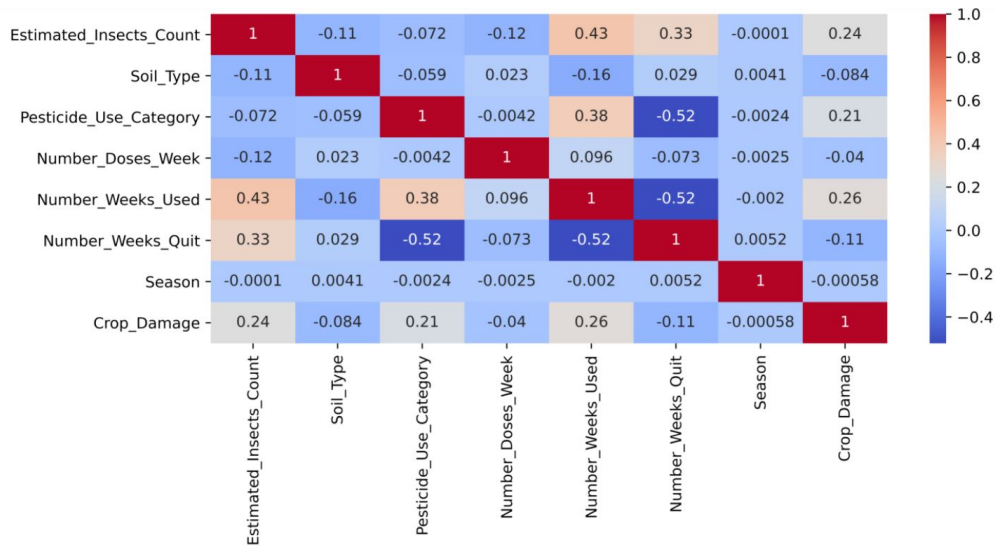


FIGURE 4.5: Heat map showing the correlation between the attributes of the crop dataset

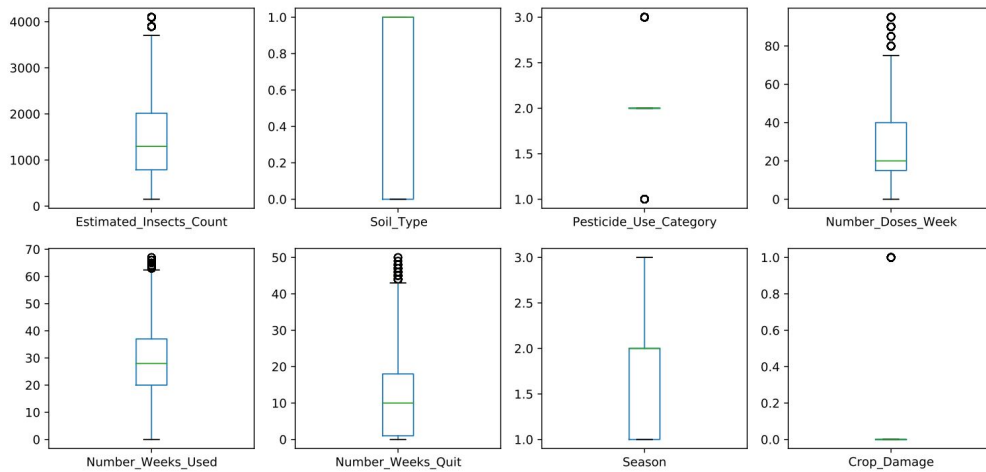


FIGURE 4.6: Attribute-based Box plots showing the presence of outliers in the crop dataset

log) per second. In a typical harvesting scenario, there are more than one harvester working simultaneously in the field. This results in overlap of the trajectories of the vehicles. A single trajectory is considered as the consolidation of all collected GPS logs belonging to a single combine harvester.

For the purpose of this study, we use the movement trajectory of a single combine harvester recorded over a period of three days. The behaviour of combine harvesters depend on the task at hand (harvesting in a field or traveling between fields). Changes in combine harvester speed, usually signal in-field or on-road actions such as turning and accelerating. During harvesting a combine harvester should operate at a maximum of 4mph in order to be efficient with minimal grain

loss and a maximum of 20mph on the road or when moving between fields [15].

The American Society of Agricultural and Biological Engineers (ASABE) recommends a typical field speed of 3mph, with a range of 2–5mph [140]. It has been shown that when the combine harvester's field speed is too high, it increases field efficiency but decreases material capacity because grain spills out on the ground [141]. Moving at high speed can result in grain loss, thus lowering the yield. The focus of this research is the identification of abnormal harvesting behaviour by focusing on local regions (i.e. local contextual anomalies). A deviation based anomaly detection approach is applied with a focus on in-field harvesting efficiency.

Trajectory mining was done using Quantum Geographic Information System (QGIS), an open-source cross-platform desktop GIS application that supports viewing, editing and visualization of geospatial data. The key focus was on the GPS logs generated in the field during harvesting, therefore visualization and map matching was done using QGIS to ensure the GPS points were mapped to a field. The second step was to extract only those points within a specific field using a bounding box. The extracted data exhibited normal harvesting behaviour with all data points below 4mph which is the maximum harvesting speed [15].

The data was further processed by removing all data points with zero speed since our interest is on GPS points associated to actual grain harvest. To create an evaluation dataset, anomalies were introduced in the original dataset by varying the vehicle speed at specific points along the trajectory such that for specific sections of the trajectory a sequential number of points had their speed increased by a random number between a given range of values above 4mph. Figure 4.7 presents the trajectory of a combine harvester after introduction of anomalies with normal points in green, anomalies in red. To visualize the effect of the introduced anomalies, outlier analysis was done using box plots (as shown in Figure 4.8), which show presence of outliers in speed and accuracy attributes.

Further data analysis was done to establish existence of any correlation between the data attributes through generation of a correlation heatmap (as shown in Figure 4.9). The map shows correlation between latitude and longitude which is expected since these are spatial coordinates. There is also a high positive correlation between the speed and class attribute which is a true representation since the anomalies introduced in the data affect the speed attribute.

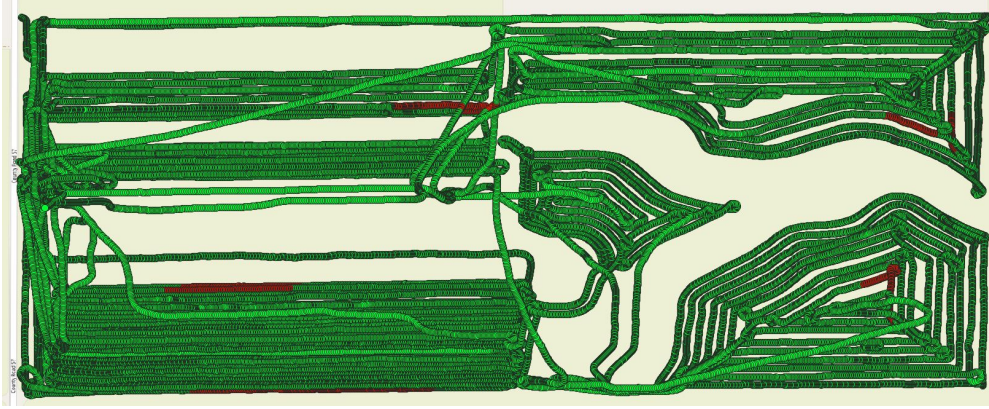


FIGURE 4.7: Field of interest: Trajectory of a combine harvester showing normal points in green and anomalies in red

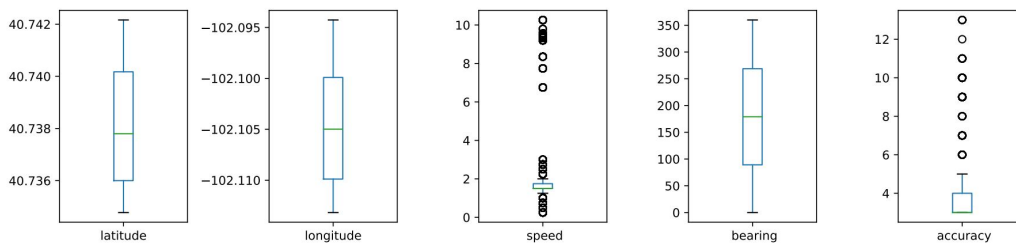


FIGURE 4.8: Attribute-based Box plots showing the presence of outliers in combine harvester GPS logs

4.4.3 Scenario C: C-ITS messages

The third level of our consideration involved anomaly detection on roads by analysing CAMs generated in a C-ITS environment. A common approach to traffic incident detection is to learn traffic patterns using previously observed accumulated traffic data and identify incidents when real-time traffic data is substantially different from the learned patterns [3], [142]. In this regard, a key feature is the speed of the vehicle, which has a direct consequence on safety, productivity and the magnitude of environmental impacts for the traffic ecosystem.

It is expected that a vehicle's speed values recorded at consecutive timestamps exhibit temporal continuity with small, calculated variations. Therefore the speed standard deviation of the set of speed of messages that belongs to a short sub-trajectory T_i^k of trajectory T_k should be low in normal condition, and could be high for anomalies. Also, speeds recorded at spatially close locations should exhibit spatial continuity with minimal variation. Detecting anomalies using CAM messages is a multivariate task that should consider a sub-trajectory size in order to be able to learn normal situations and to detect abnormal ones. Anomalies are

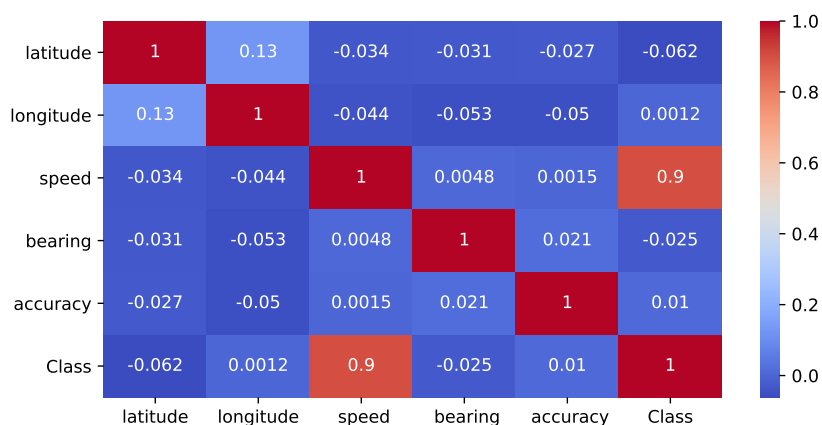


FIGURE 4.9: Heat map showing the correlation between the attributes of the combine harvester GPS logs dataset

identified as points that don't seem to fit in with the rest of the dataset, based on the premise that the majority of the occurrences in the dataset are normal.

We propose techniques based on hypothesis testing on the occurrence of an incident on a road segment. The primary assumption of our analysis is that, "if there is a traffic incident on the road, the vehicle speed, as captured from the sent messages, will substantially vary from the typical speed or the expected speed at that section". Therefore, the key hypothesis is that: *"If vehicles change their speed rapidly at a particular point, then it implies an incident has occurred"*. Our focus is on nonrecurring traffic disruptions whose occurrence is usually unexpected and random. In a scenario where an incident blocks the whole lane, we expect a substantial change in speed and also a change in heading as the vehicles encounter the incident location (as shown in figure 4.10). The main concern is on local contextual anomalies whereby an incident will affect a subtrajectory such that the contextual anomaly can be identified from the behaviour attributes, speed and heading.

A real dataset of CAMs collected from 80 vehicles in France between September

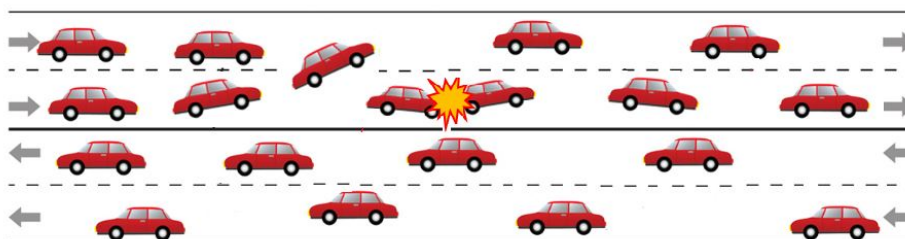


FIGURE 4.10: Illustration of an accident scenario on the road

2018 and August 2019 under a C-ITS project was available. However, the number of messages exchanged by vehicles on a road segment was not enough to build a real dataset for robust anomaly detection. Therefore, in our study we use a dataset of CAMs generated by the OMNET simulator together with SUMO, and artery plug-in. The simulation process is explained in chapter 3 section 3.4. In order to produce realistic data, the generation is calibrated according to the ETSI standard [51] specification. A vehicle sends CAMs to its neighbourhood using V2V or V2I communications. The frequency of CAM message generation varies from 10Hz to 1Hz (100 milliseconds to 1000 milliseconds).

To extract the sample data to be used in the study, trajectory mining was performed using PostgreSQL database system with the spatial extension PostGIS used for storing and processing spatial data. The focus was on the CAMs generated by vehicles which passed through Boulevard Dauphinot (route N51) in the city of Reims, France. The first step was to convert the latitude and longitude values for each message to a geometry data type using Spatial Reference Systems SRID 4326 (WGS 84), the projection for Europe. Using the *ST_DWithin* function in PostGIS to create a bounding box, all messages within a 150 meter radius of an identified reference point were extracted. This yielded messages for a 300 metre stretch of the road which forms our study area. The 300-metre limit was chosen due to the fact that C-ITS vehicles have a 150 meters communication range [143], thus a Road Side Unit (RSU) can be placed at the centre of the selected route. The extracted dataset includes trajectories for 40 vehicles.

To create an evaluation dataset, anomalies were introduced in the extracted dataset by introducing an "obstacle" at the centre of the road section [6]. When a message is in range of the obstacle (based on distance computation using latitude and longitude), the driver is said to have detected the obstacle, and the data in the message is updated. The speed is reduced by a random number between a given range of values. The following messages, in range of the obstacle according to the position of the obstacle and length of the obstacle, will be modified as well. The anomalies in the data represented incidences on the road where vehicles will have to reduce their speed as they approached the obstacle', maintain the reduced speed for the length of the obstacle and resume regular speed after passing the obstacle. The introduced obstacles mimics road incidents (like stalled vehicle, accident or debris on the road) allowing us to flag the messages of the vehicles affected by this incident as anomalous. The remaining messages were left intact. Figure 4.11 presents the trajectories on Boulevard Dauphinot (route N51) and the obstacle section showing normal points in green with anomalies in red.

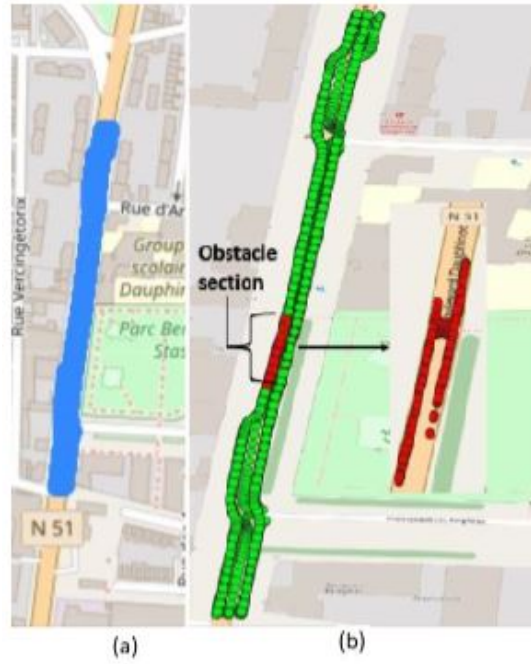


FIGURE 4.11: Area of interest; (a) trajectories on Boulevard Dauphinot (route N51), (b) Obstacle section: Normal instances in green, Anomalies in red

Multivariate anomaly detection is performed by considering the vehicle Id, timestamp, latitude, longitude, vehicle speed and heading. The data is modelled as a data stream and applied to ensemble based technique for incident detection using windowing concept.

The interest is in non-recurrent incidents, thus the anomalies of interest are resulting in an imbalanced dataset. Therefore, let \mathcal{A}_m be the set of CAM message anomalies, then:

$$\forall \text{ CAM message } m, Pr(m \in \mathcal{A}_m) \ll 1 - Pr(m \in \mathcal{A}_m).$$

And, let \mathcal{A}_T , the set of (sub-)trajectory anomalies, then:

$$\forall \text{ trajectory } T_k, Pr(T_k \in \mathcal{A}_T) \ll 1 - Pr(T_k \in \mathcal{A}_T).$$

4.5 Experimental Evaluation and Results

4.5.1 Scenario A: Crop damage

The pre-processed data was then applied to ELSCP using Kendall correlation where the aim was to evaluate its performance by looking at the true positive rate and false positive rate using a ROC curve. The AUC-ROC was also computed

as an evaluation measure of the model. In the unsupervised outlier detection setting, it is often problematic to judge the effectiveness of the algorithms in a rigorous way especially when it outputs an outlier score which is converted to a label based on a threshold. If the threshold selection is too restrictive (to minimize the number of declared outliers), then the algorithm will miss true outlier points (false negatives). On the other hand, too many false positives will be generated if the algorithm declares too many data points as outliers. To deal with this issue we performed threshold selection with the aim of identifying the threshold on the ROC Curve that results in the best balance between the true positive rate and the false positive rate for our imbalanced dataset. The Geometric Mean or G-Mean was used to compute the optimal threshold such that the best threshold was the one with the largest G-mean value.

ELSCP was trained with 67% of the data and prediction performance tested with 33% of the data. The baseline area under the ROC curve is usually set at a value of 0.5 which represents a random guess scenario where the detector has no discrimination of the normal and anomalous data. Figure 4.12 represents the ROC curve showing the best threshold value for ELSCP model. ELSCP achieved 0.641 AUC-ROC, G-mean value of 0.605 and a best threshold value of 0.1404 which shows that it is able to detect anomalies resulting in crop damage. We also evaluated ELSCP performance in terms of precision and recall. The optimal threshold for the Precision-Recall curve which focuses on the performance of a model on the positive (minority class/anomalies) was computed by optimizing the F-measure. The baseline AUCPR for the dataset was 0.17 (based on a sample size of 63589 with 10811 true positive anomalies). Based on the results obtained from precision recall curve, ELSCP achieved AUCPR of 0.277 with $F1$ score of 0.343 and best threshold value of 0.1404 (as shown in Figure 4.13). The obtained AUCPR value is acceptable since it is above the acceptable baseline for the crop dataset.

The second phase of analysis sought to determine if the detected anomalies were linked to the state of the crop being damaged at the end of the harvest season. We extracted all the data that was labelled as anomalies by the model and compared to the actual ground truth. The model detected 6324 samples as anomalies of which 1922 (30%) are actual anomalies which are linked to crop damage. In the dataset the crop state at the end of the harvest can be categorised as alive, damaged due to other causes or damaged due to pesticides. Based on the 30% actual anomalies, we categorized them according to the cause of damage where we established that 1641 samples (85.38%) were damages caused by other

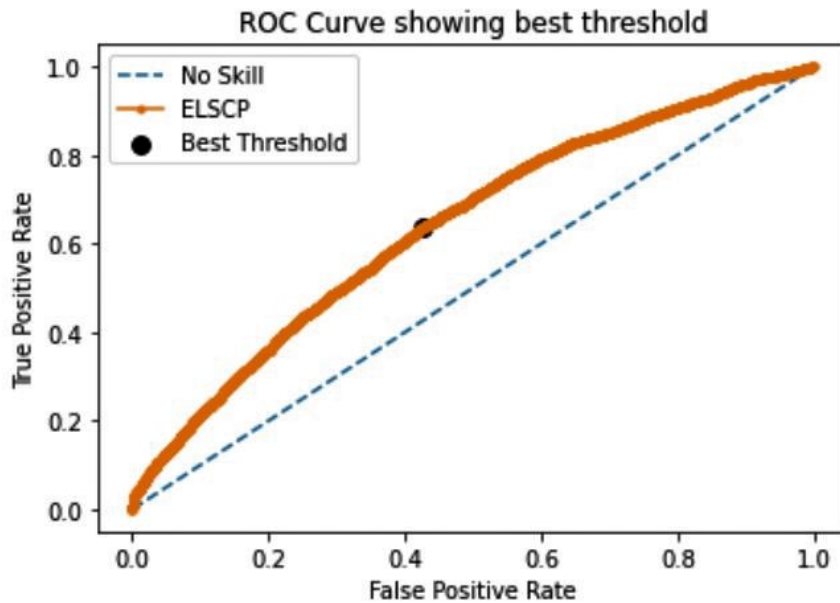


FIGURE 4.12: ELSCP ROC Curve indicating the best prediction threshold for crop dataset

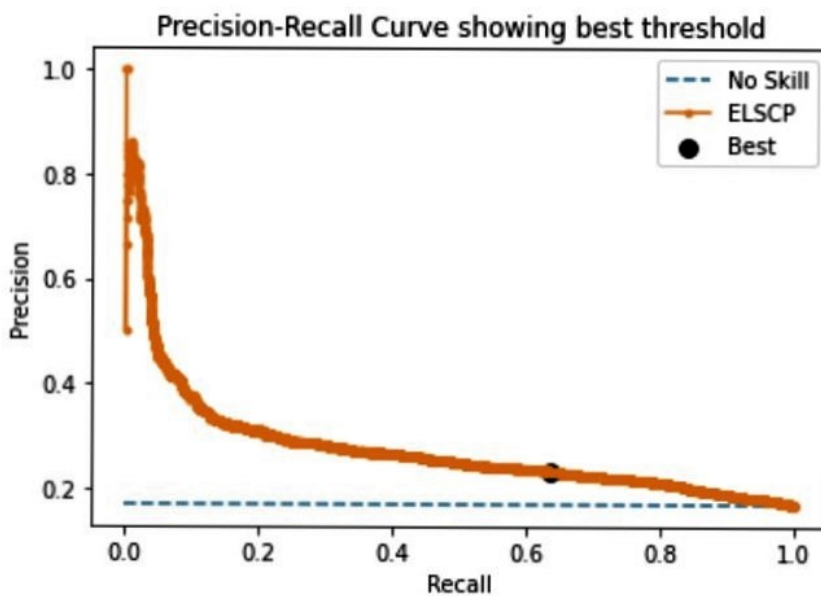


FIGURE 4.13: ELSCP Precision-Recall Curve indicating the best prediction threshold for crop dataset

sources and 281 samples (14.62%) were damages caused as a result of pesticide use.

The third phase sought to identify the anomaly detection algorithm which best estimates the anomalous instances within the crop dataset. We applied the data to ELSCP, baseline LSCP with variants of LOF as base detectors, LODA, COPOD, OCSVM and CBLOF. The performance evaluation was done by considering each

TABLE 4.2: Performance comparison of various detectors on Crop dataset

Model	AUC-ROC	AUCPR	F1 score
ELSCP	0.641	0.277	0.343
OCSVM	0.595	0.253	0.291
LODA	0.580	0.200	0.122
COPOD	0.675	0.297	0.282
CBLOF	0.636	0.226	0.212
LSCP	0.452	0.169	0.135

detectors performance in terms of AUC-ROC, AUCPR and $F1$ score. The results obtained from the experiments are summarized in Table 4.2. According to the results, we remark that our approach obtains competitive results in this context where it comes second to COPOD with an AUC-ROC margin of 5.4% (64.1% vs. 67.5%) and an AUCPR of 27.7% vs. 29.7% of COPOD which gives a margin of 3.0%. It achieves the highest $F1$ score which is 5.2% (34.3% vs. 29.1%) better than the second best detector OCSVM. Therefore, our approach is clearly very competitive in this context, and can help farmers in real-time to detect potential issues on crop health.

4.5.2 Scenario B: Combine harvester GPS data

The pre-processed data was then applied to ELSCP using Kendall correlation where the aim was to evaluate its performance by looking at the true positive rate and false positive rate using a ROC curve. The AUC-ROC was also computed as an evaluation measure of the model. The baseline AUC-ROC is usually set at a value of 0.5 which suggests that the detector has no discrimination of the normal and anomalous data, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding [144]. ELSCP was trained with 67% of the data and prediction performance tested with 33% of the data. ELSCP achieved global AUC-ROC value of 0.998 which is outstanding performance (as shown in Figure 4.14). We also evaluated ELSCP performance in terms of precision and recall. The baseline AUCPR for the dataset was 0.0199 (based on a sample size of 51580 with 1026 true positive anomalies). Based on the results obtained ELSCP achieved global AUCPR of 0.972 which is very good performance (as shown in Figure 4.15).

In the second set of experiments we sought to identify the anomaly detection algorithm which best estimates the anomalous instances within GPS logs. We

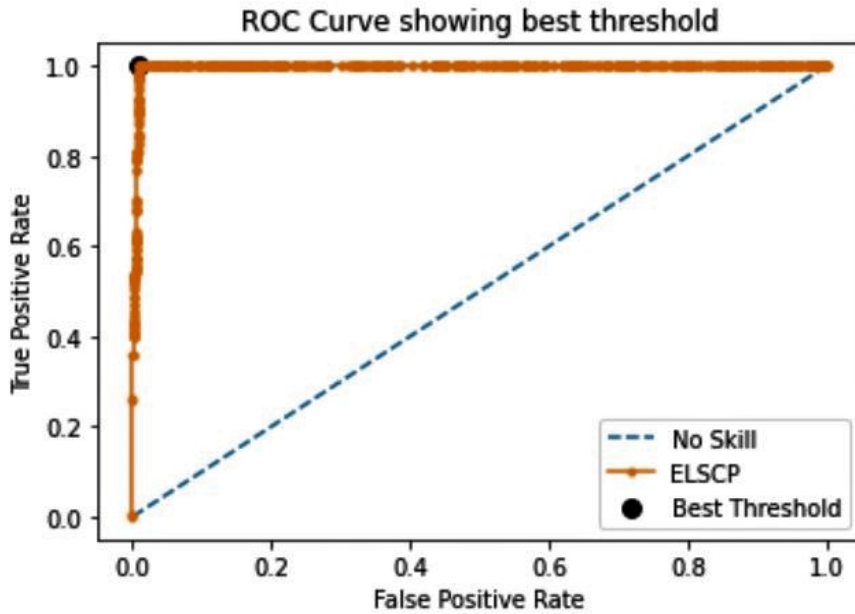


FIGURE 4.14: ELSCP ROC Curves performance evaluation for Combine harvester GPS dataset

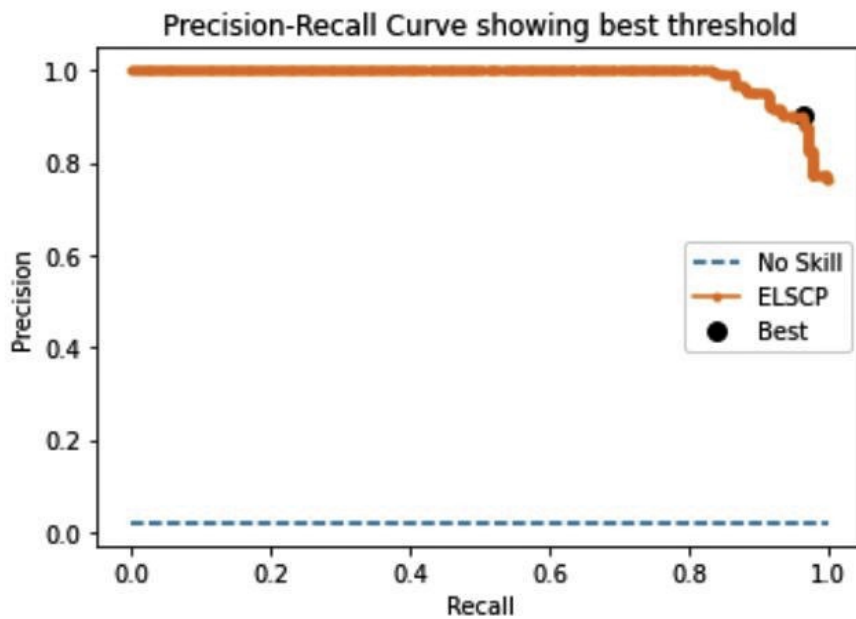


FIGURE 4.15: ELSCP Precision-Recall performance evaluation Combine harvester GPS dataset

applied the data to ELSCP, baseline LSCP with variants of LOF as base detectors, LODA, COPOD, OCSVM and CBLOF. The performance evaluation was done by considering each detectors performance in terms of AUC-ROC, AUCPR and $F1$ score. The results obtained from the experiments are summarized in Table 4.3. According to the results, we remark that our approach obtains the best scores in

TABLE 4.3: Performance comparison of various detectors on combine harvester GPS Logs

Model	AUC-ROC	AUCPR	F1 score
ELSCP	0.998	0.972	0.921
OCSVM	0.897	0.385	0.167
LODA	0.913	0.215	0.078
COPOD	0.934	0.173	0.228
CBLOF	0.756	0.038	0.014
LSCP	0.533	0.022	0.032

this context with an AUC-ROC 6.4% better than the second approach COPOD (99.8% vs. 93.4%); an AUCPR of 97.2% when the best scores of the other approaches (the one of OCSVM) is only 38.5%. The $F1$ score reflects the same thing (92.1% vs. 22.8%). Therefore, our approach is clearly the best one in this context, and can help farmers in real-time during the harvest to detect potential issues.

4.5.3 Scenario C: C-ITS CAM data

In this section, we present results obtained by experimenting anomaly detection techniques on CAMs. The algorithms were implemented in Python programming language using PySAD framework [145] which allows us to integrate batch processing algorithms from PyOD [146] and to apply them to streaming data using sliding windows. The first set of experiments were to evaluate whether the proposed ELSCP using Pearson correlation achieves better performance than LSCP. We sought to find out the model which best estimates anomalous instances and the impact of variation in window size on model performance. Multiple experiments with variation in the window size were carried out. Two base detectors HBOS and LOF were implemented in both algorithms. Table 4.4 summarises the parameters used in the experiments.

Table 4.5 summarises the results from the experiments. Based on the results it

TABLE 4.4: Experimental parameters

Parameters	Values
Window sizes	50, 100, 200, 300, 400, 500, 600
Sliding window size	50
Initial window (training set)	1000

TABLE 4.5: ELSCP AUC-ROC and AUCPR performance results for window size variation

Window size	AUC-ROC		AUCPR	
	LSCP	ELSCP	LSCP	ELSCP
50	0.7408	0.7794	0.1793	0.2157
100	0.7763	0.8065	0.2126	0.2496
200	0.8203	0.8453	0.2525	0.3125
300	0.8706	0.8977	0.2989	0.3841
400	0.8833	0.9138	0.3145	0.4208
500	0.8910	0.9247	0.3206	0.4331
600	0.8917	0.9277	0.3206	0.4373
Average	0.8392	0.8707	0.2713	0.3504

is clear that the enhancements on ELSCP yielded better performance in AUC-ROC and AUCPR. Figure 4.16 and 4.17 shows that the performance of both algorithms in terms of AUC-ROC and AUCPR increases progressively with increase in window size. In Figure 4.18, the execution time of both algorithms increases with increase in window side with LSCP giving a better time. This performance was expected since LSCP uses KDTree to compute local regions while ELSCP uses Ball Tree which is less efficient in terms of time.

The second set of experiments were to evaluate the performance of ELSCP by looking at the true positive rate and false positive rate using a ROC curve. Also,

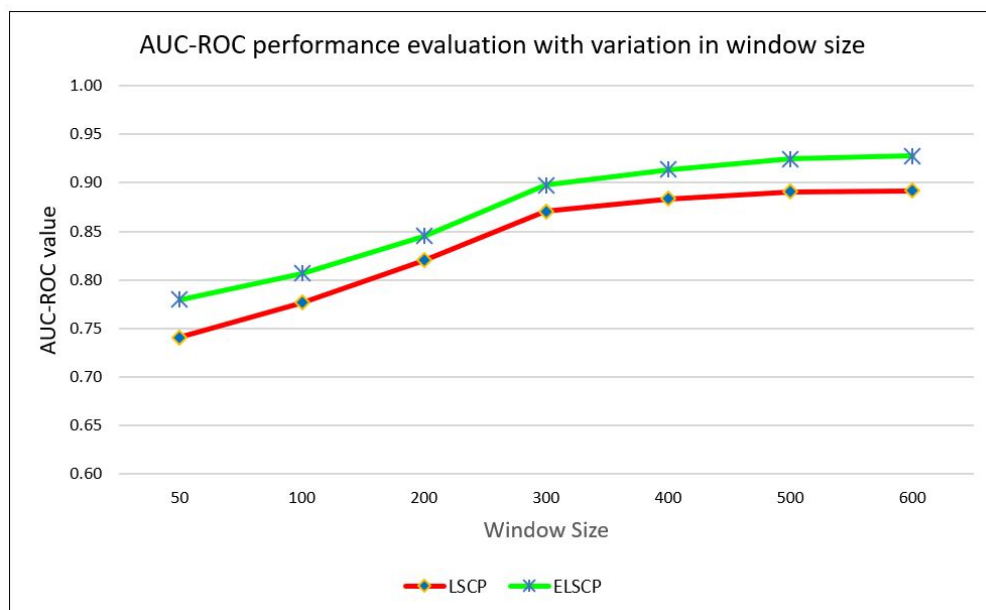


FIGURE 4.16: Comparison of models' AUC-ROC performance

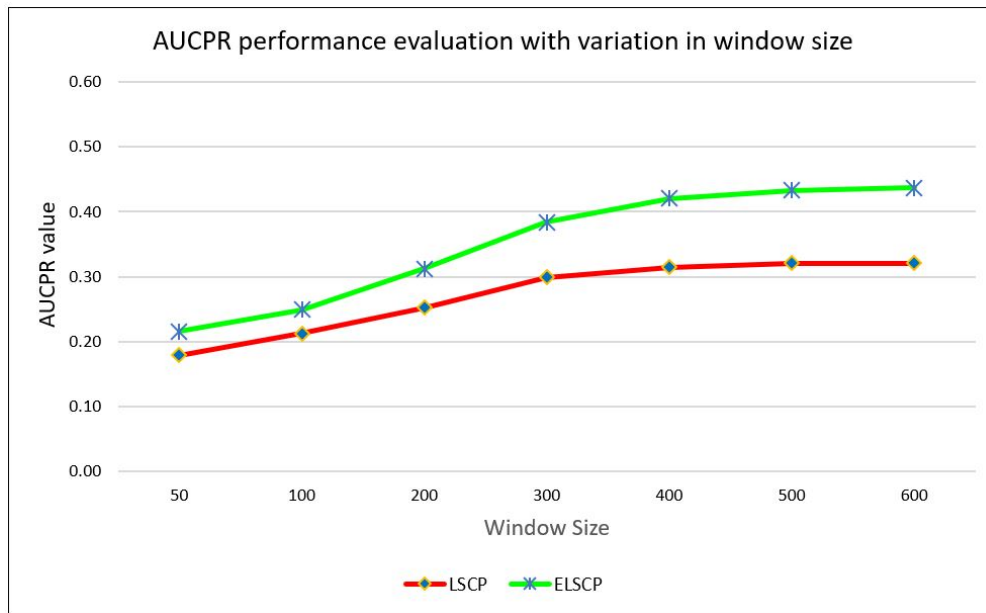


FIGURE 4.17: Comparison of models' AUCPR performance

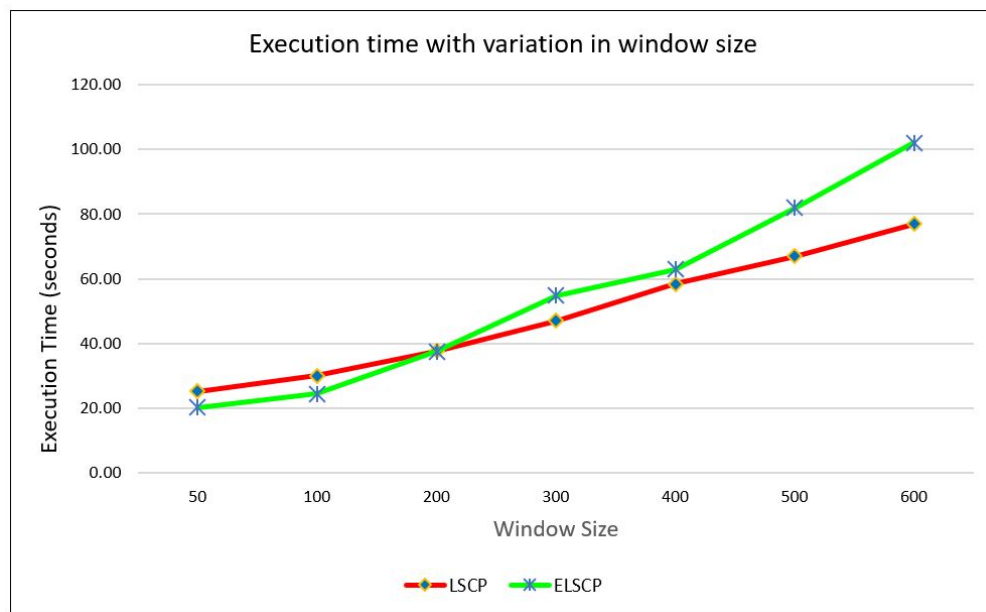


FIGURE 4.18: Comparison of models' Execution time performance

the performance in terms of precision and recall was evaluated. The baseline AUC-ROC is usually set at a value of 0.5 which suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding [144]. ELSCP was trained with 67% of the data and prediction performance tested with 33% of the data. ELSP achieved 0.97 area under the ROC curve in both the true normal data(class 0) and true anomalies(class 1) which is outstanding performance (as shown in Figure 4.19). The baseline AUCPR for the dataset was 0.087 (based on a sample size of 6341 with

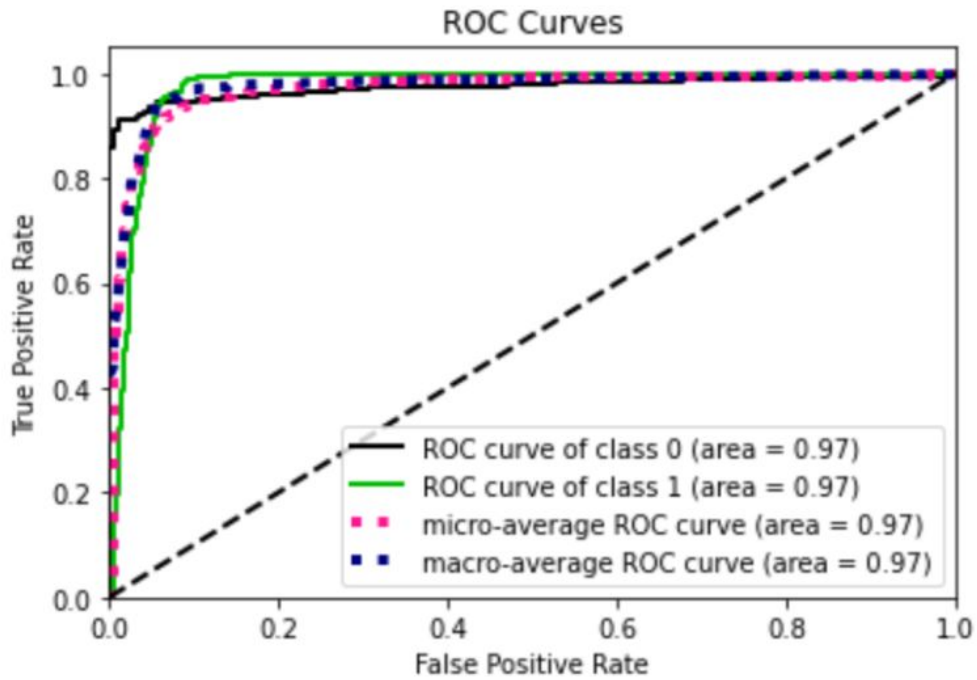


FIGURE 4.19: ELSCP ROC Curves performance evaluation

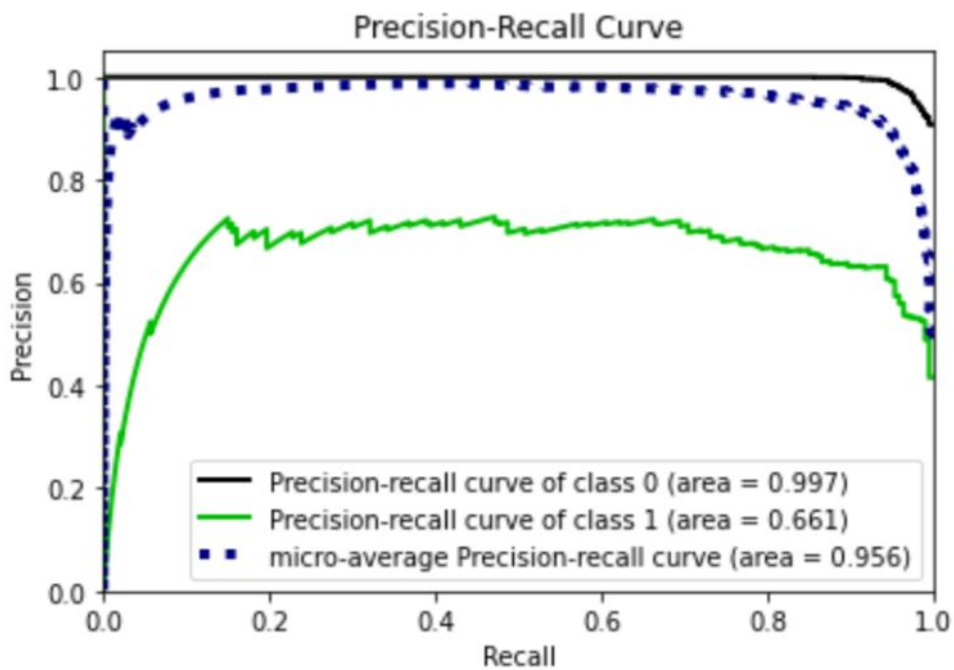


FIGURE 4.20: ELSCP Precision-Recall performance evaluation

552 true positive anomalies). Based on the results obtained from precision recall curve, ELSCP achieved 0.66 for the positive class, which is good performance.

In the third set of experiments, we sought to identify the anomaly detection algorithm which best estimates the anomalous instances within C-ITS data. We applied the data to MCD, IForest, RRCF, exact-STORM, LSCP and ELSCP

TABLE 4.6: Average AUC-ROC and AUCPR performance results for various anomaly detection models on C-ITS data

Model	AUC-ROC	AUCPR
MCD	0.7587	0.1665
IForest	0.7825	0.2210
LSCP	0.8581	0.2858
ELSCP	0.8945	0.3841
RRCF	0.6042	0.1146
Exact-STORM	0.5000	0.0885

algorithms and evaluated their performance in terms of AUC-ROC and AUCPR. Table 4.6 summarises the results from the experiments. Based on the results, it is clear that the batch processing models (MCD, IForest, LSCP and ELSCP) which have been adapted to streaming context using reference window model achieve better performance than the steaming models (RRCF and exact-STORM). In terms of AUC-ROC, both LSCP and ELSCP achieved excellent performance with ELSCP outperforming LSCP by 3.64 percent. MCD and IForest achieve acceptable performance. Exact-STORM is at the baseline and does not really seem to be discriminating between the positive and negative class. In terms of AUCPR, all models except Exact-STORM are able to discriminate between the positive and negative class since they achieve an average AUCPR value above the baseline 0.087.

4.6 Discussion

Improving agricultural productivity is critical for increasing farm profitability and fulfilling the world’s fast expanding food demand, which is fueled by rapid population increase. Precision farming seeks to reduce waste and negative outcomes by precisely focusing agricultural inputs. It has benefited from the integration of high-accuracy GPS technology into farm machinery. Several factors such as water availability, soil fertility, pest protection, timely pesticide application, and nature, all contribute to a healthy yield. Precise application of pesticides protects the crop from pests while wrong dosage can result in crop damage or even death.

A significant amount of development has been achieved in the area of anomaly detection systems, with numerous techniques proposed to address the issue of anomaly identification. Since the data is typically unlabeled and the outliers signify novel or unknown behaviours, semi-supervised or unsupervised learning

approaches are frequently utilized. Furthermore, because most of the world's data is streaming, time series data, anomaly detection algorithms must be able to analyze in real-time, learn, and make predictions. Anomaly detection's application areas are likewise highly diverse, necessitating a dependable and accurate solution. Anomaly detection approaches are domain specific, with no one-size-fits-all solution.

The most versatile, but also the most error-prone, anomaly detection technique is unsupervised anomaly detection. In this case, no data assumptions are made, i.e. the data comprises both normal and anomalous records, and algorithms should be able to distinguish between the two without any prior training. Unsupervised learning is highly preferred for real life application especially in anomaly detection since in this scenarios there is a lot of data without labels.

In this study we have performed unsupervised anomaly detection in data streams obtained from movement tracks of combine harvesters during wheat harvest, data for crop damage recorded by farmers over a period of three harvest seasons and CAM data from C-ITS environment. Based on our results, it is possible to link anomalies extracted from multivariate analysis of various features to damaged crop state at the end of harvest. Also, it is worth mentioning that deviant combine harvester behaviour can be effectively detected using our methodology. Therefore, anomaly detection could be integrated in the decision process of farm operators to improve harvesting efficiency and crop health.

In the C-ITS context, we considered anomalies which could have consequences such as an accident/incident or stalled vehicle on the road which forces the vehicles on that particular section of the road to reduce their speed substantially as they encounter the incident point. We have used streaming approaches since in real life C-ITS environments the anomaly detection would be implemented in the road side units which collect a lot of messages from the vehicles within range. This means that it would have to process the messages on the fly due to many reasons (memory limitation, response time, etc.). The implementation of the windowing concept facilitates the detection of anomalies on the fly. The merging of the techniques in the ensemble is done such that the overall complexity does not exceed that of the individual model with the highest complexity.

In the next chapter we present our contribution on behaviour analysis where we generate speed signatures from CAM messages collected in a naturalistic driving environment.

Chapter 5

Data Signatures from IoT data

5.1 Introduction

Despite the numerous advantages of highway transportation, it has certain drawbacks, such as congestion, noise, air pollution, delays, and collisions. Traffic engineers and transportation planners require information about road traffic in order to develop solutions to these problems and propose operational frameworks for the development of efficient and sustainable transportation systems. This necessitates the need for traffic studies, to systematically collect data and analyze it to understand and suggest improvement of road traffic engineering activities. Spot speed study is one such study, with the objective of estimating the distribution of speeds of cars in a stream of traffic at a specific site on a highway [147].

Speed is one of the most critical aspects affecting a driver's safety, with research demonstrating that increased speed leads to increase in collision frequencies and their severity [148]. Driving speed is regarded as a critical risk factor in road collision occurrences since drivers are frequently observed driving beyond the appropriate speed limits [149]. Drivers' speed selection is known to be impacted by a variety of factors related to driver attributes (like perseverance, negative urgency and normlessness) and the road conditions (like curves, barriers and delineators) [150]. Perseverance refers to a person's capacity to stay focused on a task that may be monotonous or challenging. Negative urgency is the propensity to behave rashly or regrettably as a result of extreme negative emotion. The notion that it is permissible to do whatever one can get away with is referred to as normlessness [150].

Understanding drivers' speed selection mechanisms and the risk variables impacting their speeding behaviour is critical for road safety. The credibility of speed

restrictions is highly related to drivers' speed choice, and it is commonly observed that drivers tend to respect speed limits when they believe them to be reasonable [151]. Underestimation of driving speed when approaching a low-traffic zone frequently results in over-speeding. As a result, drivers' compliance with speed restrictions is heavily reliant on their views of safe speed in relation to the road environment in which they are traveling [149].

Cooperative Intelligent Transport Systems(C-ITS) focus on improving safety, comfort, traffic and energy efficiency. Vehicle speed and other speed based indicators are commonly used parameters in traffic research for generation of driving profiles. Speed characteristics are significant in assessing traffic performance, examining highway consistency and safety, establishing appropriate traffic control devices and speed restrictions, and developing simulation programs [4].

The main goal for studying speed profiles is to gain a better understanding on why drivers respond in certain ways to road/traffic conditions and to discover factors which affect their actions. The speed of the vehicle has a direct consequence on safety, productivity and a magnitude of environmental impact on the traffic ecosystem. As an example, very aggressive accelerations result in higher emissions, impacting the environment negatively.

The information acquired from analysis of vehicle speeds can be useful in identification of black-spots (accident prone locations) and for gaining a better understanding of travel time. Further, it can be used to evaluate the need for infrastructure, for example speed bumps, humps and speed cameras on certain sections of the road. It is also possible to evaluate the impact of the introduction of these infrastructures by analyzing the speed profiles of vehicles as they pass through this sections and comparing it to their behaviour before the infrastructure was introduced on the road.

Naturalistic driving behaviour in real world situations can assist fleet managers and usage-based insurance providers with important information about drivers' habits, styles, and driving patterns. In this study we use a real dataset of CAMs collected from a collaborative project between road operators, car manufacturers and academia in a naturalistic driving C-ITS environment. The aim is to generate speed signatures from CAMs using segmentation and statistical data analysis approach. Speed signatures are movement patterns extracted from vehicle trajectories which can also be considered as speed profiles when analysing individual driver behaviour.

Spot speed studies are commonly used to measure the speed distribution of a traffic stream on a specific site of a motorway [152]. Our proposed approach varies

with the state of the art in that instead of using spot speeds we analyse continuous speed signatures generated within contiguous road segments. We consider the collective movement of vehicles along a particular road segment and evaluate their aggregate driving behaviour through analysis of speed signatures.

We seek to investigate the following questions:

1. How can data-driven segmentation be applied in generation of speed signatures?
2. Based on the speed signatures, what is the observed evolution of the driving behaviour on the road?
3. Do drivers maintain the speed limit restriction in force on the road segment?
4. What is the most preferred speed range by the drivers?

We make the following contributions:

1. We propose a segmentation and statistical based methodology for generation and analysis of speed signatures;
2. We evaluate the methodology using a real dataset of CAM messages generated in C-ITS naturalistic driving environment.

5.2 Problem Statement and Methodology

5.2.1 Problem Statement

Vehicles in a C-ITS environment exchange a lot of messages. Every message contains the position and movement information of the vehicle for the case of CAMs and the prevailing road condition and hazard information in the case of DENMs. The high frequency of generation of CAMs makes it possible to extract a concise mobility pattern of a vehicle. However, due to the current low penetration rate of C-ITS equipped vehicles [2] there is a challenge of insufficient amount of data for a more detailed analysis of driving behaviour on a microscopic level (route level).

We consider a situation where mobility data is captured for a specific road section for a period of eight months. On each day, a single trajectory for one or two vehicles was captured. Cases of missing data were reported in some days. In addition, the trajectories were collected at different times of the day. With

this kind of data, hourly or daily mobility analysis might not give comprehensive results since we have very few trajectories at varying times of the day.

The generation of simulated data is generally used to solve the data insufficiency issue but this still raises another issue, how to generate realistic data such that the mobility patterns are indistinguishable from those of real vehicles. In this study we analyze CAM messages with the purpose of generating collective road level speed signatures through the application of a segmentation approach and statistical analysis techniques. We consider the collective movement of vehicles along a particular road segment and evaluate their aggregate driving behaviour through analysis of the speed signatures.

5.2.2 Methodology

In a C-ITS environment, cooperative awareness is achieved through exchange of CAMs which contain the vehicle's information. The purpose of CAMs is to give dynamic information about the vehicle (i.e. speed, position, heading (direction of motion with regard to true north) etc.). A vehicle sends CAMs to its neighbourhood using V2V or V2I communications. The sending and receiving of mobility data is done through On-Board Units (OBUs) installed in the vehicles, road side units among other techniques. These data contains details that explain the movement of vehicles. Each trajectory of a vehicle is considered a multi-attribute, time-ordered sequence of locations traversed and is viewed as a trip.

A real dataset of Cooperative Awareness Messages (CAM) collected in France between September 2018 and August 2019 under the SCOOP@F project [153], [12] is used. SCOOP@F is a C-ITS project whose aim was to deploy Cooperative ITS in a nation-wide scale in France. This was to be done by equipping 2,000 km of road with Road Side Units and 3,000 vehicles with On-Board Units to enable communication between the infrastructure (the road operator) and the user (the driver). The first deployment was done in five sites: Ile de France, Bordeaux and its ring, Paris-Strasbourg highway, Brittany and Isère counties. These sites are distinguished by a wide range of road types (highways, urban expressways, two-way interurban and local roads). Its objective was to improve the safety of road users and road operating staff. It was also to assist in improving traffic management and multimodality. This was a collaborative project between road operators, car manufacturers and academia. The C-ITS CAM messages were generated by vehicles under naturalistic driving environment. The frequency of CAM message generation varied between 10Hz to 1Hz (100 milliseconds to 1000 milliseconds).

Each message has an identifier (stationid) associated with the transmitting vehicle but this vehicle is unknown. The message also includes a timestamp, latitude, longitude, altitude, speed, heading angle and the drive direction of the vehicle. Latitude, longitude, altitude will be considered as the position variables. The speed, heading angle and drive direction are used as variables of the behaviour of the transmitting vehicle. Trajectory mining was done using the open source PostgreSQL database management system, with the spatial extension PostGIS used for storing and processing spatial data.

The aim is to generate speed signatures by modeling speed variations from the start point to end point of route nationale N118 which stretches from Sèvres to Les Ulis in France. Trajectories from vehicles which travelled this route were considered. In order to ensure that all the vehicles were positioned on the correct road segment, we performed map matching using PostGIS. We started by converting the latitude and longitude values for each message to a geometry data type using Spatial Reference Systems SRID 4326 (WGS 84) which is the projection for Europe. This allowed us to view the graphical output of our queries directly in pgAdmin (PostgreSQL's graphical user-interface) against an OpenStreetMap background map using PostGIS geometry viewer. Route N118 has both an upstream and downstream section, therefore to identify the trajectories on each section we performed map matching for a few points of each trajectory. Based on the direction indications on OSM background map we were able to label each trajectory as either upstream or downstream.

The entry point for the study area is the section connecting N118 from route La Francillienne, while the exit point is the section connecting to Grande Rue as shown in figure 5.1. All the trajectories entering N118 through La Francillienne are considered to be moving in the upstream direction. The trajectories moving along N118 in the opposite direction, exiting to La Francillienne are considered downstream trajectories. A total of 32 trajectories were considered with 14 upstream and 18 downstream trajectories. Considering the fact that the messages are highly sampled with frequency of up to 10 points per second, we opted to follow an aggregated approach for speed signature generation.

The first step in the segmentation process was to group the messages based on equal length segments from a specific reference point such that each message is in a unique road segment. We chose the first reference point to be the geometry of the first message at the entry point of route N118. Then using the *ST_DWithin* function in PostGIS to create a bounding box, we extracted all messages within a 1 kilometre radius of the reference message, this group forms the first segment



FIGURE 5.1: Area of interest; (a) trajectories on the full stretch of route N118, (b) Entry point, (c) Exit point

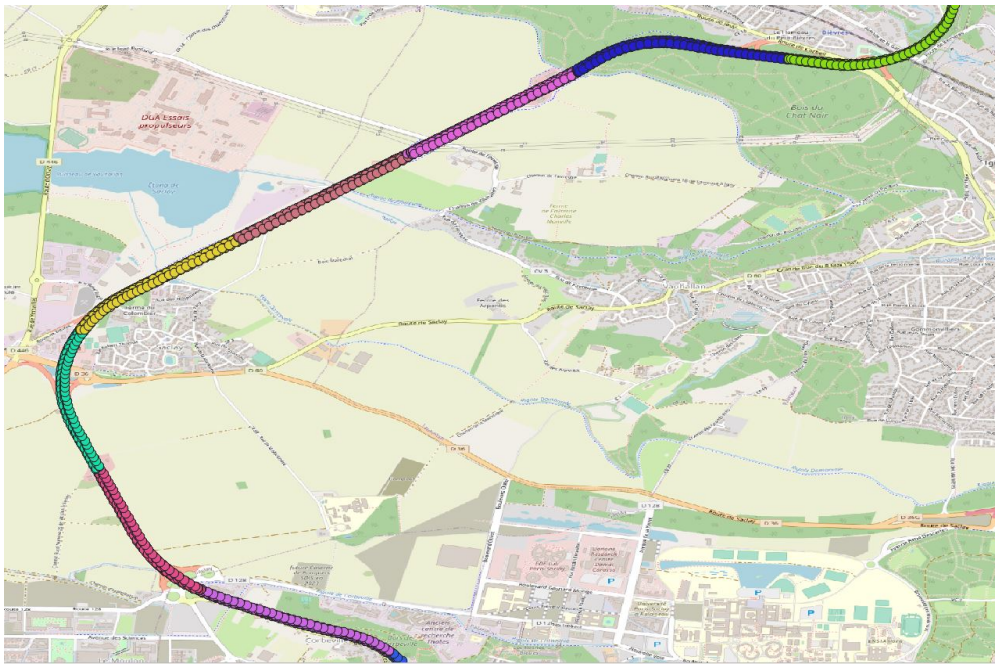


FIGURE 5.2: Sampled road section showing the extracted one kilometre length segments

and is extracted and stored. The second segment is generated from the remaining messages using the same process by moving the reference point to the first message of the unprocessed points at the edge of the previous bounding box. This process iterates until we have all the messages grouped in 1 kilometre length segments. Figure 5.2 presents a sampled section of the road showing the one kilometre length trajectory segments with each segment represented by a unique colour.

Generation and interpretation of speed signatures is achieved by evaluating the data using descriptive statistical analysis techniques. The approaches of interest are measures of relative standing, measures of central tendency and measures of variability, which are techniques for transportation data analysis [154]. Measures of relative standing such as percentile rank and quartiles provide information

about the relative placements of observations within a dataset. Percentile rank is essential when measuring driving speeds on a road segment. The speed of a driver is compared to the speeds of all other drivers that drove on the road segment during the measuring period.

Speed percentiles are methods for determining effective and appropriate speed restrictions. The 50th and 85th percentiles are the two most significant speed percentiles to consider. The median speed of the observed dataset is represented by the 50th percentile. This percentile shows the speed at which half of the observed cars are below and half are above. The average speed of the traffic flow is represented by the 50th percentile of speed.

Definition 5.1: *85th Percentile Speed (V_{85}):* The 85th percentile speed is the speed at or below which 85 percent of the free-flowing vehicles travel. This implies that if the 85th percentile speed is 90 km/h, then 85 percent of the observed drivers were driving at speeds less than 90 km/h and 15 percent were driving at speeds more than 90 km/h.

Quartiles are percentage values that divide data into quarters. One-quarter of the data is filtered out at the first quartile (25th percentile). The cutoff for half of the data is the second quartile (50th percentile), commonly known as the median. The point above which one-quarter of the data is located is the third quartile (75th percentile). Measures of central tendency, such as median and mean, represent the "*centrality*" of an observation within a dataset [154].

Definition 5.2: *Mean Speed:* The mean speed is the most common measure of central tendency which is calculated by summing all the measured speeds and dividing by the sample size. When analyzing the collective mean speed for all the instances in a dataset, the population mean (μ) is computed as defined in (Equation 5.1).

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (5.1)$$

where N is the total number of instances in the dataset.

In the case where a sample of the data is used, for example when computing the mean speed for data within a particular road segment, the arithmetic / sample mean is computed as defined in (Equation 5.2).

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (5.2)$$

where n is the total number of instances in the sample (road segment).

Definition 5.3: *The Median (50th Percentile) Speed* is an alternate statistic to

the mean or average speed for describing the speed distribution's central tendency. It is the speed attained or surpassed by precisely half of the cars. It is resistant to the influence of extreme observations or outliers and is independent of the shape of the data. Despite the popularity of the mean, the median is a more trustworthy metric in cases where the data contains a large number of outlying observations.

Definition 5.4: *Speed Dispersion:* The speed dispersion refers to the normal spread in vehicle speeds observed in a study section/segment.

The measures of variability, interquartile range (IQR) and standard deviation (SD), characterize the spread or distribution of data around the mean or any other measure of central tendency. The IQR is a numerical ratio that represents the difference between the first and third quartiles. It is also resistant to outliers in the data. The standard deviation, calculated as the square root of the variance, describes the data dispersion around the mean. When analyzing the collective standard deviation of speed for all the instances in a dataset, the population standard deviation is computed as defined in (Equation 5.3).

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (5.3)$$

where N is the total number of instances in the dataset.

To compute the standard deviation of speed for data within a particular road segment, sample standard deviation is computed as defined in (Equation 5.4).

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (5.4)$$

where n is the total number of instances in the sample (road segment).

5.3 Experimental Evaluation and Results

Statistical data analysis was done using R software. Using the extracted segments we generated aggregate statistics for speed where we computed the minimum, maximum, mean, median, standard deviation and 85th percentile for both upstream and downstream sections of the road.

TABLE 5.1: Summary of vehicle speed

Measure	<i>Upstream</i>	<i>Downstream</i>
Minimum	0.00	0.00
1st Quartile	16.60	20.92
Median	50.58	65.09
Mean	50.28	57.48
3rd Quartile	82.33	88.63
Maximum	123.59	131.00
Standard deviation	34.02	35.49
IQR	65.73	67.67
85th percentile	91.66	95.62
Speed limit	90.00	80.00

5.3.1 Evaluation of collective speed signatures

Table 5.1 presents a summary of vehicle speed aggregates in kilometers per hour along route N118.

Based on these speed aggregates, we evaluated the variation in the aggregate measures and compared against the speed limits on the road ¹, 90km/h on the upstream sections and 80km/h in the downstream section. Table 5.2 presents the summary speed characteristics per road segment for the upstream segments of route N118. Table 5.3 presents the summary speed characteristics per road segment for the downstream segments of route N118. The results for upstream and downstream speed signatures are shown in figure 5.3 and figure 5.4 respectively. The variation on median speed per segment is shown in figure 5.5

Based on the analysis of the upstream section, the average speed varies slightly above the 90km/h limit for segments 2 to 7 and segment 9. Then there is a decline from segment 10 with the most notable change being at the 16th segment where the average speed is 15.88 km/h. The median speed also varies above the 90km/h limit for segments 2 to 9, then it starts declining with the lowest value of 7.85 km/h at segment 16. This behaviour is also confirmed by the trend of the standard deviation which progressively increases from the 7th to the 8th segment, followed by a decline on the 9th segments and a sharp increase to 45.14 on the 10th segment. This may be attributed to the curvature of the road (as shown in figure 5.6), an incident or existence of traffic calming infrastructure. On the downstream section, the average speed is more steady and under the speed limit. The median speed is also below the speed limit except for segments 12, 16, 18 and

¹[https://truck-simulator.fandom.com/wiki/N118_\(France\)](https://truck-simulator.fandom.com/wiki/N118_(France))

TABLE 5.2: Upstream summary speed characteristics per road segment

Segment	Min	Q_1	Median	Mean	Q_3	Max	SD	V_{85}
1	62.46	79.22	87.32	86.08	94.43	102.82	10.22	96.19
2	75.71	89.35	95.11	94.78	101.56	107.06	7.49	103.10
3	76.28	90.30	92.83	92.98	97.02	105.19	5.32	98.02
4	83.74	89.51	93.74	95.35	100.41	113.33	7.21	102.03
5	76.90	86.18	90.29	90.69	93.85	110.16	6.90	96.01
6	85.75	93.82	98.21	100.53	109.76	111.60	7.82	109.84
7	57.10	94.99	102.06	98.30	105.73	114.01	12.8	109.73
8	6.88	43.34	90.65	74.71	95.58	104.76	29.70	97.88
9	46.12	78.21	98.75	92.46	106.15	112.82	16.16	107.10
10	0.00	15.30	47.12	58.67	106.40	123.59	45.14	109.67
11	0.00	26.75	60.12	59.91	97.60	113.26	36.68	101.60
12	0.00	12.98	36.18	46.41	83.45	110.70	35.68	87.59
13	0.00	28.48	40.64	49.37	83.20	102.49	29.04	89.97
14	0.00	16.77	69.37	56.60	86.16	109.01	32.50	88.95
15	6.12	22.43	68.47	59.52	96.91	112.18	38.70	102.71
16	0.00	4.36	7.85	15.88	14.76	110.05	23.61	23.42
17	0.00	6.55	13.46	22.14	32.47	97.70	23.83	40.03
18	0.00	25.51	50.90	47.49	65.47	95.83	26.66	75.76
19	6.98	37.59	63.83	58.18	78.36	96.95	23.50	82.48
20	13.50	36.29	55.06	54.77	74.62	98.32	21.85	80.46
21	14.94	50.89	60.70	58.03	69.05	90.29	16.93	74.24
22	9.83	45.60	51.53	50.21	57.74	73.55	11.45	60.70

19 which are slightly above the speed limit. Based on the analysis of the median speed variation we can conclude that for most parts of the road in both upstream and downstream, the speed limit policy is maintained.

5.3.2 Evaluation of speed dispersion and preferred speed ranges

In addition to this evaluation, we compare individual vehicle speed signatures by generating a box-and-whisker plot (boxplot), which depicts the drivers' speed range patterns. The boxplot depicts the central tendency of the driver speed signature, as well as the IQR and the number of drivers that violated the speed limit. The median, first and third quartiles, as well as the maximum and minimum of the speed signatures, were calculated prior to producing the plot. The computed Upstream values are presented in the table 5.4 while table 5.5 presents the downstream values.

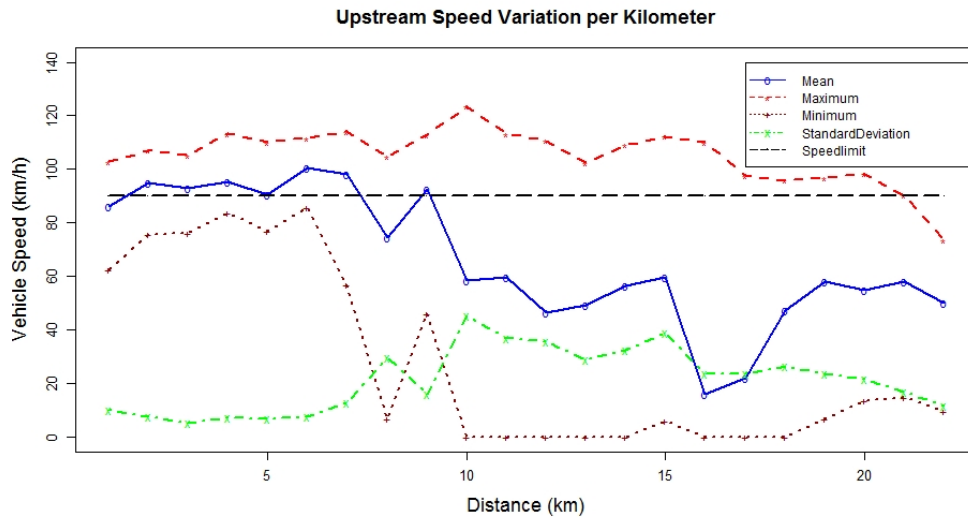


FIGURE 5.3: Upstream Speed signature variation.

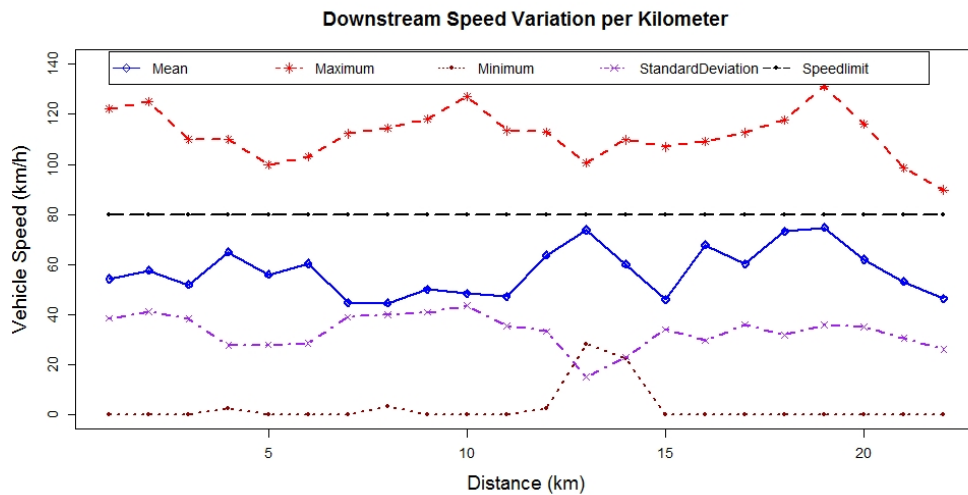


FIGURE 5.4: Downstream speed signature variation.

The IQR, which represents the speed range each driver most frequently traveled within, may be determined from the box plots as shown in figure 5.7 and figure 5.8. The box part of the graph shows the speed range between the first Quartile (25th percentile) and the third Quartile (75th percentile), while lines within each box show the median value. The whiskers extend to the most extreme data points not considered outliers, while the outliers are plotted individually using 'o' symbol. The boxes with their whiskers show the full spectrum of the drivers speed from minimum to maximum and the dashed red line represents the speed limit (90 km/h on the upstream lanes and 80 km/h on the downstream lanes).

Based on the results depicted in figure 5.7 for the upstream section of the road, six drivers (D01, D46, D53, D61, D80 and D96) exclusively operate within

TABLE 5.3: Downstream summary speed characteristics per road segment

Segment	Min	Q_1	Median	Mean	Q_3	Max	SD	V_{85}
1	0.00	16.67	38.16	54.26	91.22	122.08	38.57	94.96
2	0.00	16.45	43.47	57.61	98.44	125.14	41.27	107.88
3	0.00	14.87	35.71	51.94	90.50	110.09	38.27	92.81
4	2.63	39.94	74.20	64.89	85.95	109.94	27.85	91.02
5	0.00	26.69	56.48	56.03	82.23	99.79	28.00	84.37
6	0.00	32.72	67.09	60.41	86.38	103.25	28.64	88.88
7	0.00	16.00	26.01	44.82	94.64	112.25	39.03	99.40
8	3.31	14.18	21.22	44.55	96.77	114.62	40.16	104.33
9	0.00	17.18	25.96	50.06	102.95	118.22	41.04	107.24
10	0.00	15.12	25.34	48.50	107.12	127.01	43.76	112.32
11	0.00	15.92	33.59	47.29	83.48	113.47	35.54	94.04
12	2.48	27.76	82.30	63.80	91.19	113.00	33.33	94.29
13	28.33	64.48	77.18	73.80	84.96	100.76	15.35	88.19
14	22.54	39.09	58.01	60.09	84.46	109.80	23.13	85.80
15	0.00	16.78	34.90	45.99	83.63	106.96	34.14	89.78
16	0.00	49.61	81.68	67.80	88.67	109.22	29.68	94.69
17	0.00	17.26	77.83	60.21	90.88	112.64	36.12	94.44
18	0.00	61.06	84.46	73.20	95.67	117.79	31.94	99.68
19	0.00	45.67	87.71	74.78	100.31	131.00	35.98	106.52
20	0.00	18.70	79.24	62.00	89.14	116.10	35.26	94.57
21	0.00	17.86	66.53	53.24	78.99	98.64	30.61	84.02
22	0.00	11.95	58.61	46.37	67.14	89.86	26.10	69.74

the speed limit set for the road. All the other drivers except D04 and D67, operate within the speed limit at least 75% of the time. Drivers D04, D06 and D067 have a substantial number of outliers. In terms of speed variability, D08 exhibits a larger variability with 50% of the values between 9.94 km/h to 85.11 km/h. D27 and D60 also have a wide variable speed distribution. It is also evident that the speed distribution of D27, D53 and D80 is skewed to the right with majority of the speed values above the median speed value. On the other hand, D06 and D61 are skewed to the left with majority of the speed values below the median speed value.

The results for the downstream section of the road depicted in figure 5.8 show that only one driver D12 operated within the speed limit. Drivers D79 and D96 maintained the speed limit more than 75% of the time. Driver D10 was exclusively above the speed limit. In terms of speed variability, D90 exhibits a larger variability with 50% of the values between 19.66 km/h to 93.92 km/h. D35 also show variability with the data skewed to the left with majority of the speed values

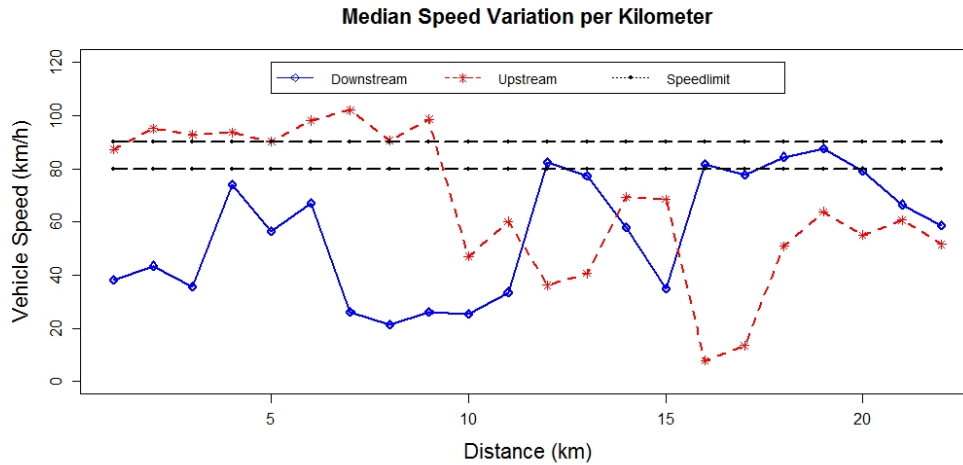


FIGURE 5.5: Median speed signature variation for both upstream and downstream sections of the road.



(A) Road segment 8



(B) Road segment 16

FIGURE 5.6: Representation of road topology for segment 8 and 16 with upstream trajectory points in blue and downstream points in purple

below the median speed value. D09, D45, D37, D70 and D90 also have their speed distribution skewed to the left. D28, D79, D81 and D98 have the speed distribution skewed to the right with majority of the speed values above the median speed value.

The speed range most frequently maintained per road segment is shown in figure 5.9 and figure 5.10. From figure 5.9, there are a number of road segments with outliers, the most affected being segment 16 on the upstream section of the road. This could be caused by the curvature of the road and the fact that

TABLE 5.4: Upstream Drivers' speed box plot parameter values

Driver	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum
D01	0.00	21.62	44.06	58.01	83.84
D04	23.69	87.81	92.30	98.23	111.60
D06	9.83	63.22	73.39	80.09	94.64
D08	0.00	9.94	49.91	85.11	115.24
D27	0.00	6.70	16.20	67.72	113.26
D46	6.98	32.11	40.25	50.22	64.40
D53	41.08	45.83	47.32	51.03	55.19
D60	0.00	15.16	35.62	65.84	114.01
D61	41.76	61.82	70.69	75.82	84.17
D67	37.84	83.57	90.50	97.95	123.59
D68	0.00	15.66	41.62	64.22	112.18
D78	0.00	24.18	39.01	55.20	97.70
D80	39.28	48.65	56.74	68.09	73.94
D96	0.00	13.00	37.12	55.01	75.64

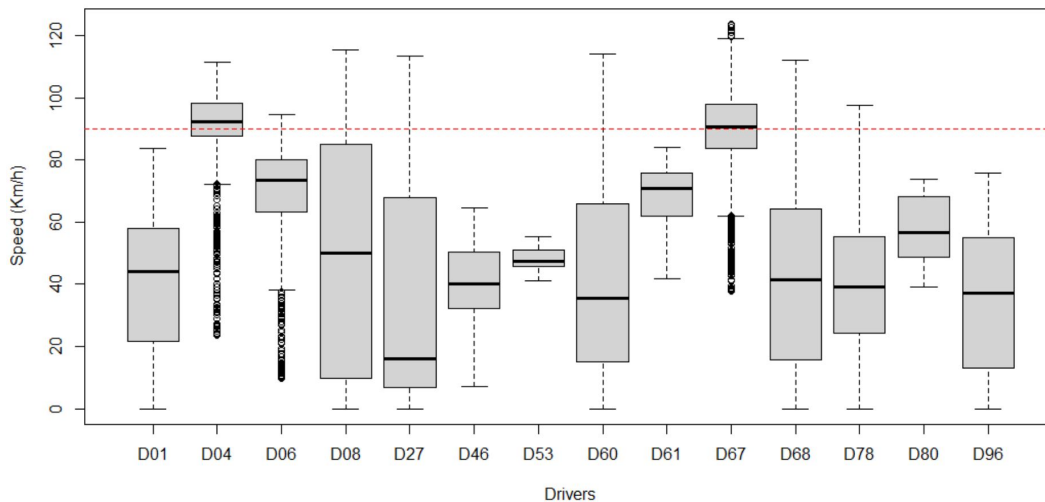


FIGURE 5.7: Upstream Box-and-whisker plots of the drivers' vehicle speed ranges.

there are arterial roads joining and exiting which implies the presence of traffic calming infrastructure. It is also evident that the data is skewed with a wide speed dispersion on the downstream section of the road, particularly segments 1-3, 7-11 and 15 which are skewed to the right.

The third set of experiments were performed with the aim of comparing the drivers' vehicle speed based on the number of times they travelled within specific speed ranges. We defined four bins for the speed in both the upstream and downstream sections of the road. The following speed range bins were chosen for the

TABLE 5.5: Downstream Drivers' speed box plot parameter values

Driver	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum
D09	40.43	61.02	76.09	82.17	90.07
D10	108.90	111.00	113.90	115.60	121.50
D12	30.56	59.94	66.35	71.39	77.40
D20	48.28	69.41	79.13	88.70	106.02
D25	74.70	95.92	100.37	104.13	113.11
D28	70.81	86.76	91.85	100.49	115.96
D35	3.13	47.38	79.67	94.00	127.01
D37	66.71	91.26	100.94	107.21	116.96
D38	46.26	70.88	83.52	96.30	103.64
D42	73.55	84.28	86.06	89.32	107.68
D45	0.00	73.19	83.12	88.34	94.68
D55	50.26	78.68	85.43	91.84	111.10
D56	27.11	81.22	85.32	91.22	125.68
D70	53.06	73.40	87.98	93.61	99.07
D79	0.00	12.20	23.26	43.63	104.72
D81	57.67	67.12	79.52	97.89	131.00
D90	0.00	19.66	71.42	93.92	123.26
D98	0.00	13.28	20.21	34.44	108.79

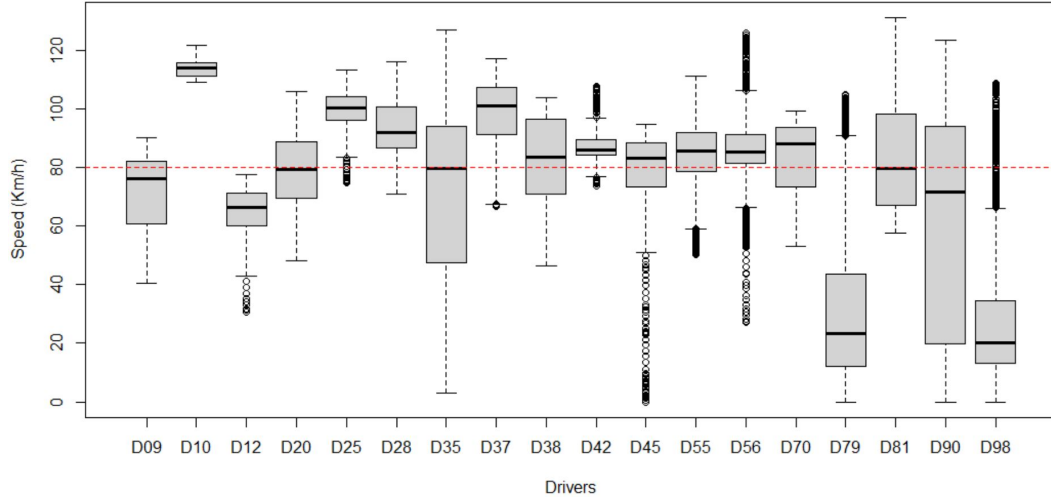


FIGURE 5.8: Downstream Box-and-whisker plots of the drivers' vehicle speed ranges.

upstream section of the road:

- Between 0 to 10.9 km/h;
- Between 11 to 50.9 km/h;
- Between 51 to 90 km/h;

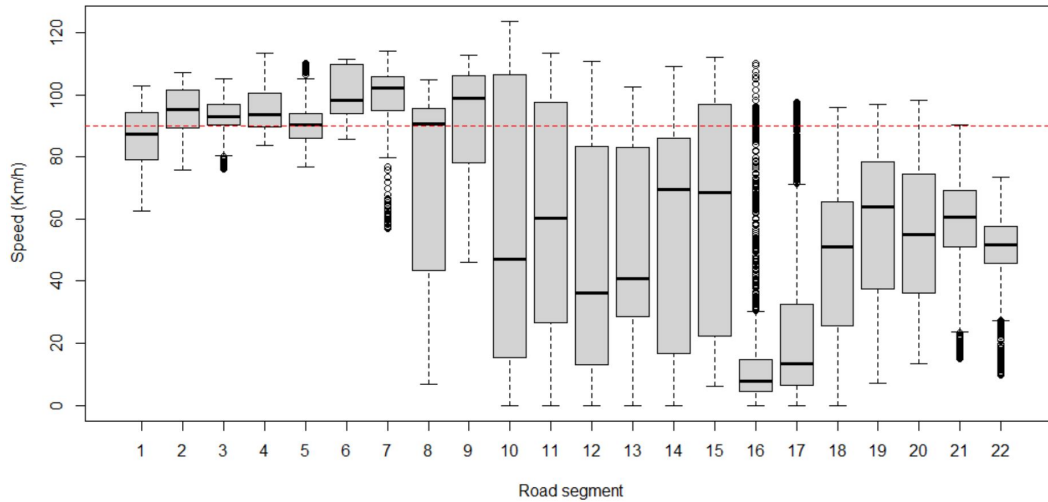


FIGURE 5.9: Upstream Box-and-whisker plots of the vehicles' speed ranges per road segment.

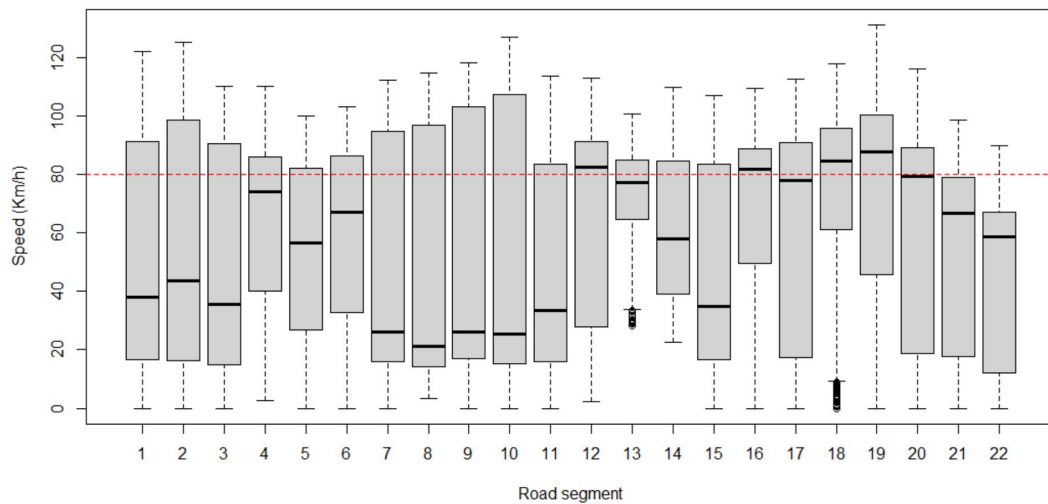


FIGURE 5.10: Downstream Box-and-whisker plots of the vehicles' speed ranges per road segment.

- Over the speed limit.

The following speed range bins were chosen for the downstream section of the road:

- Between 0 to 10.9 km/h;
- Between 11 to 50.9 km/h;
- Between 51 to 80 km/h;
- Over the speed limit.

To provide an overall view of each driver's tendency to drive in the four zones, a percentage stacked bar chart format was generated. The data is visualized in

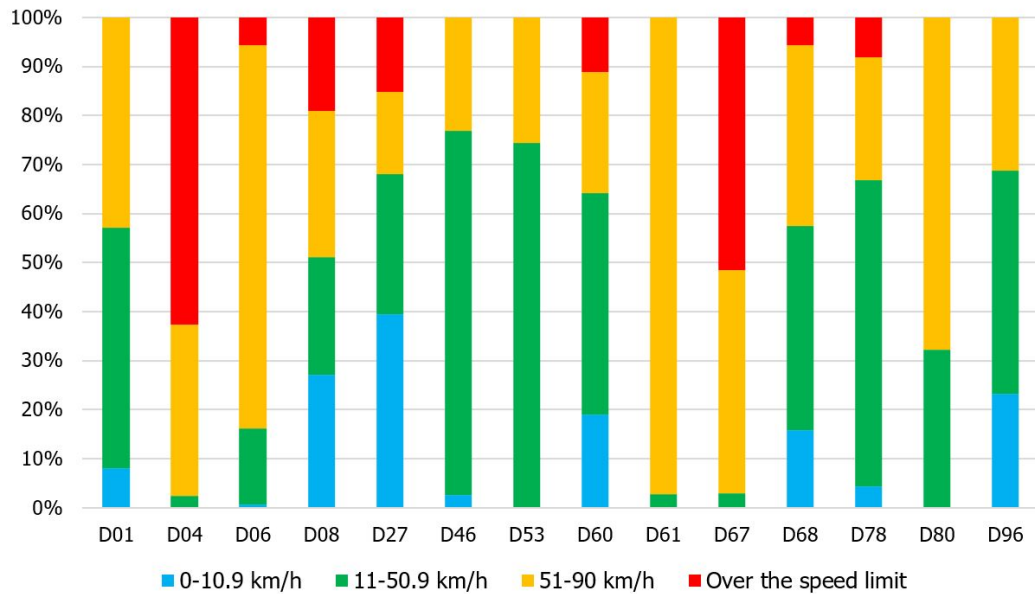


FIGURE 5.11: Upstream Stacked bar chart of the drivers' chosen range of vehicle speeds.

this way to provide a quick perspective of each driver's behaviour. Based on the results for the upstream section (presented in figure 5.11), except for D04 and D61, all the other drivers drive within the speed limit 80% of the time. Nine drivers (63%) prefer driving below 51 km/h half the time. It is clear that the most preferred speed range is 11-80 km/h.

Based on the results for the downstream section (presented in figure 5.12), it is clear that except for driver D12, all the other drivers drove over the speed limit at some point. Driver D10 was exclusively over the speed limit with D25, D28, D37 and D42 driving over the speed limit more than 90% of the time. With the majority of the drivers driving over the speed limit, it might be necessary to review the speed limit or introduce variable speed limit so as to mitigate the displayed behaviour.

5.4 Discussion

Most nations employ speed limits regulations to govern the speed of vehicles. Many research studies have focused on the influence that speed restrictions have on drivers' choice of speed since it is one of the most important components in understanding the link between speed and safety. Speed limits are significant because they decrease the disparities in vehicle speeds between drivers on the same route, which improves safety. Setting speed limits aims to strike a good

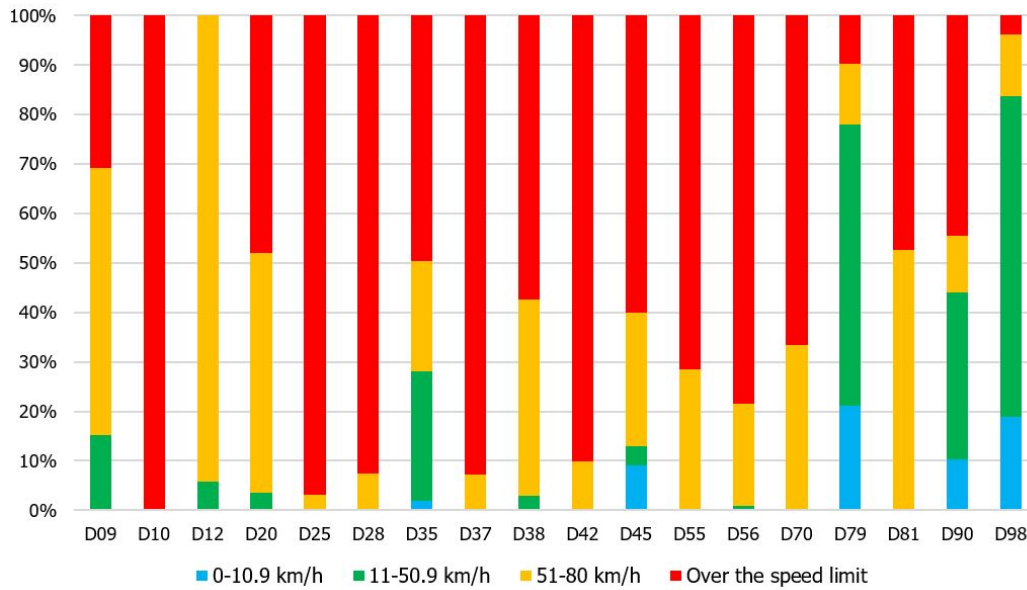


FIGURE 5.12: Downstream Stacked bar chart of the drivers' chosen range of vehicle speeds.

balance between safety and mobility for a given road.

Individual drivers' speed choices are the result of a complex interaction of numerous factors, including the highway's design speed, the danger of collecting speeding fines (and the penalties thereof), the published speed limit, and the anticipated likelihood of a crash [155]. If the geometric aspects of the road encourage them to go faster, drivers fail to realise their travel speeds and drive faster than the published speed limits. In real-world traffic, some drivers will opt to obey the speed limit signs, while others may believe that the speed limit is too low and will pick their own pace. Variable speed limits are thought to have a better effect on speed compliance since they may be varied based on road conditions [150].

In this work we considered the analysis of speed signatures generated from C-ITS messages with the aim of understanding driving behaviour evolution under a naturalistic driving environment. C-ITS data is used to assess the appropriateness of the listed speed limits on the route under consideration. We have shown that with the application of segmentation and aggregate statistics, one is able to get a better understanding of general driving behaviour and also infer information that relates to the road condition and traffic situation. The findings of the studies suggest that vehicle speed is important when evaluating driving performance and identifying differences in driving behaviour.

A possible application area for the generated speed signatures is in Autonomous vehicles where the signatures can be used as inputs in order to check if the driving

is following normal speed/expected naturalistic speed behaviour. The signatures can also be used by traffic engineers and transportation planners to evaluate the effectiveness of the posted speed limits such that adjustments can be done based on the observed behaviour of the drivers. This may necessitate a reduction or an increase in the current speed limit by evaluating the 85th percentile speed of the traffic data stream.

With the current uptake of C-ITS there remains a challenge of insufficient amount of data for a more detailed analysis of driving profiles on a microscopic level (route level). It will be of interest to do a more detailed analysis of the data to identify factors causing or influencing the observed behaviour, especially on the spike points. Further, useful knowledge can be gained by considering the individual behaviour of the vehicles on these particular road segments. This can be done through outlier detection in order to identify incident points or unique behaviour among certain drivers.

The DENM messages can also be used to extract points where incidents were reported and then use these locations as POIs for traffic incident detection. By extracting the speed signatures of road segments, it could now be possible to discover whether there is a change in driver behaviours over a road segment on a time period (and therefore discover road incidents) by comparing them with the road segment speed signature using anomaly detection method such as Isolation Forest (IForest), Angle-Based Outlier Detector (ABOD), K-Nearest Neighbours (KNN) etc.[146].

In the next chapter we present our contribution on trajectory linking where the idea is to link identifiers to the vehicles which generated them. The real dataset applied during signature generation and analysis in this chapter will also be utilized for data linking in the next chapter.

Chapter 6

IoT data linking

6.1 Introduction

The development of wireless communication technology, geographical information systems, embedded positioning devices and ubiquitous devices has facilitated collection and storage of vast quantities of IoT mobility data. The collection of mobility data is done by online or offline means through devices attached/carried by the moving objects, road side units among other techniques. These data typically contains information describing the movement of people, goods, vehicles, aircraft, animals, natural phenomena (hurricanes, tornadoes, and ocean currents), etc. Each trace of a moving entity is a multi-attribute, time-ordered sequence of locations.

When considering moving objects especially on road networks, the paths taken are linked to the prevailing traffic environment and conditions. When analyzing these trajectories it is important to incorporate the environmental information so as to gain a better understanding on the movement patterns [156]. The behavioural and lifestyle aspects of a moving object can be discovered from the examination of its daily trajectories. Trajectory pattern analysis is invaluable in applications such as: recommender systems, public security systems, and path planning in emergency evacuations [90].

Moving objects may have different strategies when they need to report their locations to a central repository, such as time-based, distance-based, and prediction-based strategies. They may also suspend the communication with a central server for a while and resume later. The overall result is that the lengths and time stamps of the trajectories will be different and the trajectories may also be segmented with gaps (missing readings). Each trip in a trajectory dataset includes an

identification (ID) of the device it was recorded from. Device IDs enables chaining of consecutive trips of the same vehicle to rebuild movement over a longer period of time, which provides a better insight into mobility patterns. However, device IDs may periodically change for privacy or some other reasons, which clearly limits the analysis. Thus, to be able to gain knowledge from trajectories a method for chaining anonymous trajectories and filling missing gaps is required.

The trajectory linking problem, which tries to link trajectories to the user who generates them, is crucial to research. We refer to a trajectory with an unlinked owner as an unlinked trajectory, and a trajectory with a clear owner as a linked trajectory. Assuming a set of unconnected trajectories and a set of users, the goal of trajectory linking is to find a mapping that connects the trajectories to their associated users. This problem is difficult because there is need to simultaneously determine the commonality of a user's trajectories while distinguishing the trajectories of various users (i.e., the intra-class similarity and inter-class difference).

Trajectory-User Linking (TUL) problem is currently an active research area in location based social networks where the aim is to identify the users who generate check-in trajectory data. In this study we look at this problem in relation to trajectories generated by vehicles on a constraint road network.

We seek to investigate the following questions:

1. How can similarity in movement patterns and background information be applied in trajectory linking?
2. Given multiple identifiers assigned to a vehicle are we able to group the identifiers and detect those which belong to the same vehicle?

We make the following contributions:

1. We propose to solve the TUL problem by chaining anonymous trajectories to potential vehicles by considering similarity in movement patterns.
2. A distance function is proposed to measure similarity between trajectories with a combination of background knowledge in extraction of linked data.
3. A trajectory mining framework is proposed and evaluated on a real dataset of CAMs generated in C-ITS naturalistic driving environment.
4. The results of data linking are validated using map matching technique.

6.2 Problem Statement and Methodology

6.2.1 Problem Statement

The vehicles of an Intelligent Transport System (ITS) exchange a lot of messages. Every message sent is generated with an identifier of the transmitting vehicle. To respect the user privacy, the identifiers of each vehicle are changed regularly. An identifier is kept only over a time interval. The issue we want to study is, given several identifiers, are we able to detect those which belong to the same vehicle? We adopt the definition of [37] for Trajectory User-Linkability problem:

Let $T_{v_i} = \{m_{i1}, m_{i2}, \dots, m_{in}\}$ denote a trajectory generated by the vehicle v_i during a time interval, where $m_{ij} (j \in [1, n])$ is a message sent from a specific location at time t_j . Given that the identifier is changed after a time period, trajectory $T_x = \{m_1, m_2, \dots, m_y\}$ generated by the same vehicle in the next time interval with a different identifier is considered unlinked. TUL can thus be defined as:

Suppose we have a number of unlinked trajectories $T = \{t_1, \dots, t_m\}$ generated by a set of vehicles $V = v_1, \dots, v_n (m \gg n)$, TUL learns a function that links unlinked trajectories to the vehicles: $T \rightarrow V$

6.2.2 Methodology

In a C-ITS environment cooperative awareness is achieved through exchange of CAMs which contain position information. This can act as a privacy threat to the drivers since an eavesdropper can be able to create a detailed mobility pattern of the driver. To mitigate this, Pseudonym schemes are used to provide anonymous communication. A pseudonym generally provides authentication for vehicles which can use multiple pseudonyms in order to guarantee unlinkability of actions [157]. This involves the change of pseudonyms after a preset time period so as to prevent linkability of one pseudonym to another which can in turn result in the identity of a vehicle and consequently that of the driver being revealed if one is able to identify the home address.

Although pseudonyms protect identification, important information in beacon communications compromises vehicle location privacy. Furthermore, because of the constant broadcast of CAM signals and the necessity for exact measurements in the messages for safety-related applications, vehicles are increasingly exposed to tracking attacks [158]. Furthermore, the vehicles travel inside the confines

of traffic laws and roadways, making it easier to estimate the target vehicle's movement characteristics.

The aim of this study is to generate continuous trajectories using anonymous data while ensuring that privacy is preserved. Our mining framework starts with **Data pre-processing** which transforms the raw trajectory data into a form that can be correctly analyzed. It entails the identification and correction of errors in the data. This is followed by feature selection where the most relevant input variables to the task at hand are identified and selected. There is also a need to do data transformation by changing the scales and/or variable distributions due to the use of different scales during data capture.

Trajectory linking: In order to link the trajectories we consider the following conditions for triggering CAM generation as specified in ETSI EN 302 637-2 standard [51] (detailed discussion in chapter 3 section 3.4):

- If the absolute difference between the current heading value of the vehicle and the heading value included in the last transmitted CAM by the same vehicle exceeds 4° ;
- If the distance between the current position of the vehicle and the position included in the last transmitted CAM by the same vehicle exceeds 4 meters;
- If the absolute difference between the current speed of the vehicle and the speed included in the last transmitted CAM by the same vehicle exceeds 0.5 m/s.

A CAM is generated every 100 milliseconds at a CAM transmission rate of 10Hz. The generated CAMS contain static data variables such as vehicle length, width, and the confidence level of heading, speed, acceleration, curvature, and yaw rate, which may be used in data linking. Furthermore, during a time frame of 100 milliseconds, the speed and geographic position change very slightly. For example, if a C-ITS vehicle with a length of 4.4 metres and a speed of 80 km/h (22.2 ms) exists. Within 100 milliseconds, it will have travelled 2.2 metres (illustrated in figure 6.1). This signifies that the geographic position of the vehicle's length overlaps by at least 50%. Therefore, no other vehicle can physically be located at the same position within that time period. This shows that a vehicle's CAMs may be linked depending on their geographic location. The linkability of CAMs may be used to map entire CAM traces of a given vehicle's journeys [118].

A first work consists in grouping as much as possible the different identifiers which represent sub-trajectories of one vehicle. A complete grouping with all the

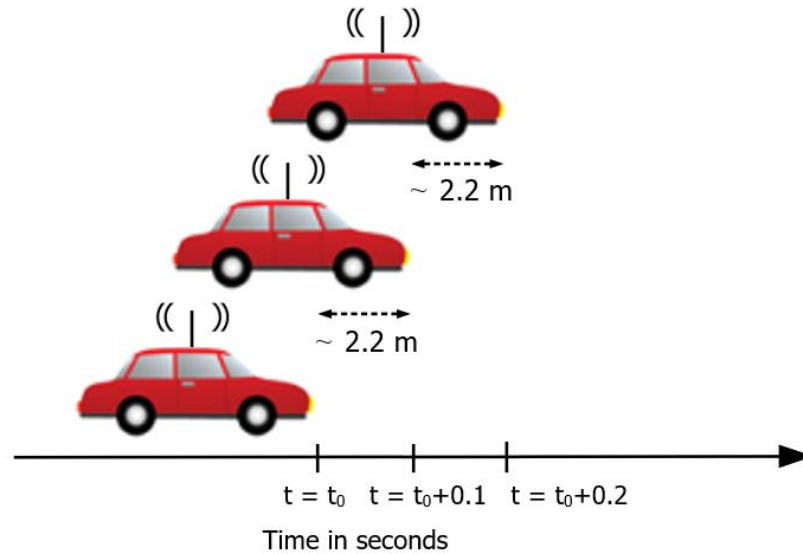


FIGURE 6.1: Representation of movement of a C-ITS vehicle within 100 milliseconds based on a speed of 80 km/h

identifiers of each vehicle may be difficult to obtain but grouping some identifiers can be obtained. For example, if the last message of an identifier is spatially and temporally close to the first message obtained with another identifier and the change in attributes like speed and heading angle is consistent, then the change of identifier from the last message to the first one is obtained for the same vehicle. Thus the two identifiers are linked and belong to the same vehicle. In this example, the work consists in defining a reliable link between two messages with different identifiers.

Then we detect the contradictions between messages. For instance, if two messages give the same localization at the same time, then their identifiers cannot belong to the same vehicle. These contradictions help to define the group of identifiers for each vehicle by rejecting the identifiers leading to a contradiction. To enhance the accuracy of trajectory linking, aspects of the vehicle's mobility need to be incorporated under **Mobility pattern mining** step. In addition to the three constraints for CAM generation, the driving direction of a vehicle is used as a parameter to filter matched identifiers where by a true match is one where the direction of motion of the trajectories is the same. Nearest neighbour (spatial) search is performed to extract the trajectory closest to the end point of the current trajectory both in space and time. We also extract the first and the last ten points per identifier for use in mobility similarity analysis.

Similarity calculation step extracts similar trajectories by computing how close they are in space, time and the similarity in movement direction. Segments

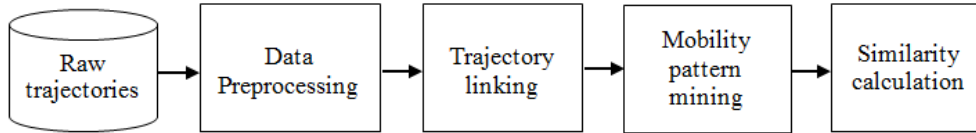


FIGURE 6.2: Trajectory mining framework.

are similar if they form a continuous sequence in space and time. The spherical distance measure is used to compute the minimum distance between two longitude/latitude geometries. Given point p_1 and p_2 , with ϕ_1, ϕ_2 as the latitude of the two points, λ_1, λ_2 as the longitude and $\Delta\lambda, \Delta\phi$ as their absolute differences. The central angle between them is defined as:

$$\Delta\sigma = \arccos(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Delta\lambda)) \quad (6.1)$$

The spherical distance d between the two points is defined as ¹:

$$d = r\Delta\sigma \quad (6.2)$$

where r is the radius of the earth(6371 km), $\Delta\sigma$ is the central angle between the two points.

The framework to be followed in the analysis is shown in figure 6.2.

6.2.3 Dataset Description

In our study we used a real dataset of Cooperative Awareness Messages (CAM) collected in France between September 2018 and August 2019 under the SCOOP@F project [153], [12]. The purpose of CAMs is to give dynamic information about the vehicle (i.e. speed, position, heading (direction of motion with regard to true north) etc.). A vehicle sends CAMs to its neighbourhood using Vehicle-to-Vehicle (V2V) or Vehicle-to-Infrastructure (V2I) communications. The frequency of CAM message generation varies from 10Hz to 1Hz (100 milliseconds to 1000 milliseconds). Each CAM is uniquely defined by a stationid (pseudonym) and timestamp.

During the SCOOP@F project, the issued pseudonyms were stored in form of pools for a specific duration, with the lifespan for vehicles being one week. Each vehicle had a maximum of ten parallel pseudonyms valid during the same time span. Each vehicle picked a new pseudonym from its pool based on a Round-Robin algorithm (as shown in figure 6.3). The Round-Robin technique ensures that the re-use of a pseudonym is not performed in the same order to try and

¹https://en.wikipedia.org/wiki/Great-circle_distance

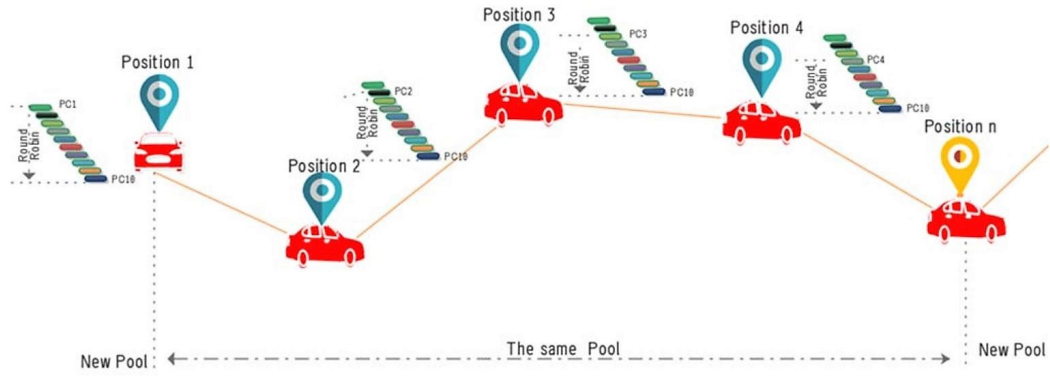


FIGURE 6.3: Illustration of the pseudonym change strategy in the SCOOP@F project [53]

prevent any attempt of tracking. The Pseudonym Change (PC) occurred after every 40 000 signatures or one hour depending on the condition attained first, and at each ignition of the vehicle. In the dataset each of the 80 vehicles was assigned unique stationids which were changed periodically resulting in a total number of 3866 unique IDs and a total number of 10,174,437 CAM messages sent.

In this study, each message has an identifier (id) associated with the transmitting vehicle but this vehicle is unknown. The message also includes a time stamp (time), the localization of the vehicle with latitude (lat), longitude (long) and altitude (alt), the speed (speed), the heading angle (angle) of the vehicle and the drive direction (direction). Thus the message is a data defined with 8 variables: id, time, lat, long, alt, speed, angle, direction. The variables lat, long, alt are the three position variables. Speed, angle and direction variables are used as variables of the behaviour of the transmitting vehicle.

6.3 Experimental Evaluation and Results

We performed trajectory mining using PostgreSQL database with the spatial extension PostGIS used for storing and processing spatial data. We also used Quantum GIS (QGIS)², an open-source cross-platform desktop geographic information system application that supports viewing, editing, and analysis of geospatial data. QGIS was majorly used for visualization and map-matching of the trajectories as a validation step. Map-matching is the process of placing each location point on a digital map, referencing it to the most relevant neighbouring road connection, and reconstructing the path between each road link to fully recreate the vehicle's or car's journey. Since our data is highly sampled with 10 points per second,

²<https://qgis.org/>



FIGURE 6.4: Distribution of all trajectories.

the projection of the data points on the base map in QGIS yielded a precise full path reconstruction of the vehicle trajectories on the roads. Data pre-processing was done by removing noise from the data. The distribution of all trajectories is shown in figure 6.4. We then extracted origin-destination pairs from the trajectories whereby an origin is the first message of each trajectory and a destination is the last message of the trajectory. Figure 6.5 shows the distribution of the origin (green colour)-destination (red colour) pairs.

Considering the fact that each vehicle was assigned multiple identifiers, we sought to link identifiers which occurred on the same day. Taking the destination points, we extracted the nearest origin point within 170 meters and also filtered out the results by implementing the CAM generation trigger conditions as additional constraints. The distance computation was done using the *ST_DistanceSpheroid* function in PostgreSQL which gives the linear distance between two longitude/latitude points. We also used the CAM generating frequency of 100 – 1000 milliseconds as a constraint in order to get exact matches in time and space.

During matching we were specifically targeting the matches which occurred on the same date within a short period of time (in few seconds) so as to get trajectories which are continuous in space and time. Also, as a test for continuity, the matched trajectories had to be traveling in the same direction during the change of identifiers. After processing all the trajectories we were able to get 867 matching/linked ID pairs with the month of April having the highest matches at

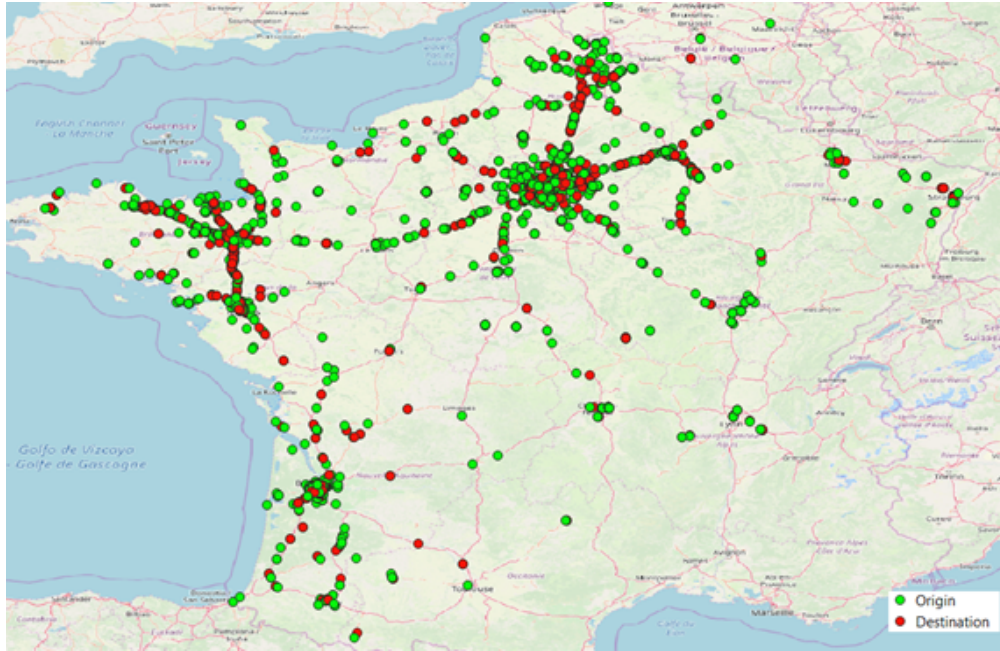


FIGURE 6.5: Distribution of origin-destination pairs.

124 IDs. We selected 45 trajectories generated on the 5th and 6th April 2019 (as shown in figure 6.6) and after processing 10 trajectories linked with others to generate a total of 35 trajectories. The highest number of linkages per trajectory was four trajectories where ID 1 linked to 2 then 3 and finally 4 both in time and space as shown in figure 6.7, thus generating one continuous trajectory as shown in figure 6.8. To validate the linkage/matching of trajectories, we performed map-matching to ensure that the trajectories are on the same road and moving in the same direction.

The fact that the trajectories are constrained by a road network increases the probability of linking trajectory segments to the generating user given background knowledge and behavioural aspects of movement like speed, heading angle and drive direction. However, complete linkage of all segments to the generating users is a difficult task and might not be possible. This is proven by the fact that out of 3866 trajectories, we were only able to link 867 pairs which is 22.43% of the total number of trajectories. The monthly analysis of linked trajectories is shown in figure 6.9 which indicates the total number of IDs per month, the total number of IDs linked per month and the linkage percentage.

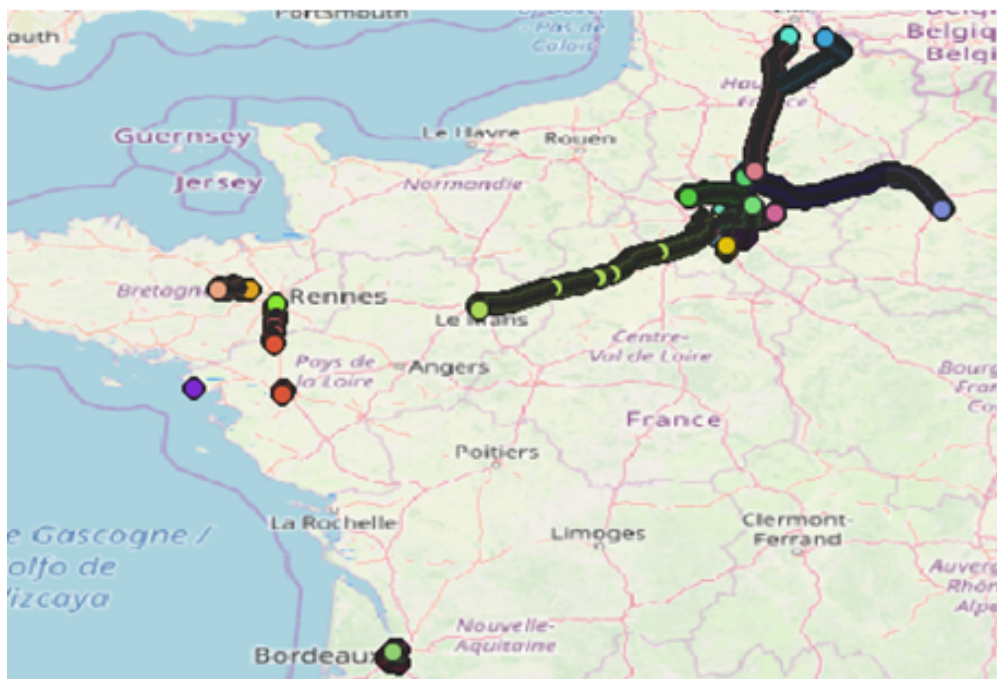


FIGURE 6.6: Distribution of trajectories for the 5th and 6th of April 2019.



FIGURE 6.7: Continuity validation of linked trajectories.

6.4 Discussion

Trajectory data is ordinarily characterized by raw collection of time ordered spatio-temporal points which capture the motion of an object in geographical space over time. Depending on the capability of the device used, these movement track data at every instant of time can include additional attributes like the speed, acceleration, direction of motion etc. The trajectories of moving objects express a concise overview of their behaviour. When analyzing the trajectory data, we do not only consider the consolidation of the recorded point data but are more concerned



FIGURE 6.8: Continuous trajectory after linking four trajectories.

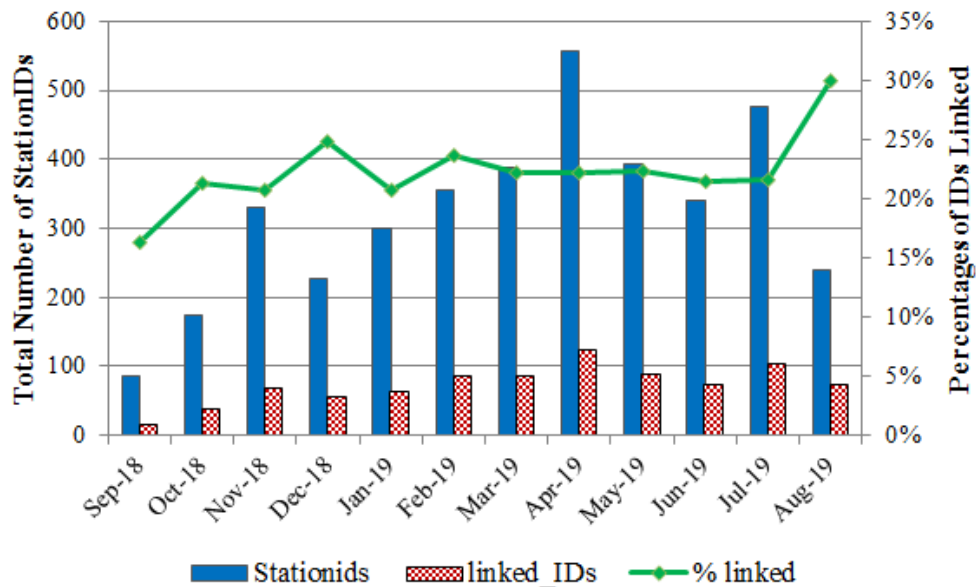


FIGURE 6.9: Monthly analysis of linked trajectories.

with extracting and understanding the semantic meaning of the trajectory.

Vehicles in an Intelligent Transport Network exchange a lot of messages. Every message sent is generated with an identifier of the transmitting vehicle. To respect the user privacy, an identifier is kept only over a specified time interval. The need that arises is, given that multiple identifiers are assigned to a vehicle, are we able to group the identifiers and detect those which belong to the same vehicle? We solved this Trajectory-User Linking problem by chaining anonymous trajectories to potential vehicles by considering similarity in movement patterns.

Our objective was to link identifiers which occurred on the same date within a short period of time (in few seconds) so as to get trajectories which are continuous in space and time. Also, as a test for continuity, the matched trajectories had to be traveling in the same direction during the change of identifiers. The fact that the trajectories are constrained by a road network increases the probability of linking trajectory segments to the generating user.

Based on our analysis, it is possible to link trajectories to the generating users if other distinguishing attributes (like speed, heading angle, altitude and drive

direction) and background knowledge on generation of the messages are considered when performing similarity analysis. It is also worth noting that the use of pseudonyms as a privacy and security measure has been proven to be a viable approach since we were not able to break the unlinkability requirement.

Chapter 7

Conclusion

In this thesis, we addressed the problem of analyzing IoT data with a focus on anomaly detection in data streams and driver behaviour analysis. A review of the state of the art was realized in chapter 2 where a detailed discussion on smart agriculture, ITS, Anomaly detection, trajectory data mining, speed profile analysis, and trajectory user linking approaches was presented. We made contributions in the contexts of smart agriculture and C-ITS, with a focus on C-ITS.

A significant amount of development has been achieved in the area of anomaly detection systems, with numerous techniques proposed to address the issue of anomaly identification. Anomaly detection's application areas are likewise highly diverse, necessitating a dependable and accurate solution. Unsupervised learning is highly preferred for real-life applications, especially in anomaly detection since there is a lot of data without labels in this scenario. In chapter 4, we proposed a new algorithm based on an ensemble anomaly detector called Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP) for stream anomaly detection. On this basis, we defined an unsupervised data-driven methodology that is applied in three case studies, with the objective of detecting anomalies on the fly.

In the first case study, we performed unsupervised anomaly detection on crop data where we studied the link between crop state (damaged or not) and detected anomalies by analysing data for crop damage recorded by farmers over a period of three harvest seasons. On the basis of our results, it was possible to link anomalies extracted from multivariate analysis of various features to damaged crop state at the end of harvest. The second study is dedicated to anomaly detection in obtained data streams from GPS tracks of combine harvesters collected during wheat harvest. On the basis of our results, deviant combine-harvester behaviour could be effectively detected using our methodology. The performance obtained

was evaluated against other approaches, and the results obtained are relevant (based on known performance metrics) with the possibility of its implementation in real time to detect anomalies and assist farmers during harvest. Therefore, anomaly detection could be integrated in the decision process of farm operators to improve harvesting efficiency and crop health.

In the third case study, C-ITS CAM messages collected from a collaborative project between road operators, car manufacturers and academia were analysed in order to detect anomalies on the road. We consider anomalies which could have consequences such as an accident/incident or stalled vehicle on the road which forces the vehicles on that particular section of the road to reduce their speed substantially as they encounter the incident point. We have used streaming approaches since in real life C-ITS environments the anomaly detection would be implemented in the road side units which collect a lot of messages from the vehicles within range. This means that it would have to process the messages on the fly due to many reasons (memory limitation, response time, etc.). Based on our results, ELSCP is able to detect anomalies from CAMs. This detection could be integrated in the decision process of road operators in order to improve safety and traffic flow. Timely detection of anomalies is critical especially for emergency response teams resulting in improved efficiency in rescue operations.

In chapter 5, we considered the analysis of speed signatures generated from C-ITS messages with the aim of understanding driving behaviour evolution under a naturalistic driving environment. We have shown that with the application of segmentation and aggregate statistics, one is able to get a better understanding of general driving behaviour and also infer information that relates to the road condition and traffic situation. With the current uptake of C-ITS there remains a challenge of insufficient amount of data for a more detailed analysis of driving profiles on a microscopic level (route level).

In chapter 6, we considered the trajectory-linking problem and applied it to messages generated by vehicles in C-ITS. Based on our analysis, it is possible to link trajectories to the generating users if other distinguishing attributes (like speed, heading angle, altitude and drive direction) and background knowledge on generation of the messages are considered when performing similarity analysis. It is also worth noting that the use of pseudonyms as a privacy and security measure has been proven to be a viable approach since we were not able to break the unlinkability requirement.

7.1 Limitations

The first limitation of ELSCP is on the extraction of neighbouring data points that constitute the local region of a test instance using distance metrics applied to KNN Ball Tree algorithm. This approach brings two challenges:

1. It takes much time to determine the nearest neighbours of the test instance;
2. The performance in a multidimensional space may be affected, especially when many features or attributes are irrelevant.

To remedy this problem, local-region definition could be solved by the use of fast approximate methods [41] or by prototyping [42], which can significantly reduce the required time to set up the local domain because not all data points are required by these techniques. The second limitation is on the generation of the pseudo-ground truth, where we applied a simple maximisation technique. This could be improved by considering exact strategies, for instance, with the active pruning of base detectors [43].

The third limitation is on the availability of C-ITS data. With the current uptake of C-ITS there remains a challenge of insufficient amount of data for a more detailed analysis of driving profiles on a microscopic level (route level). With the current data it is not possible to effectively analyse the data for hourly or weekly analysis of driving behaviour on specific roads.

7.2 Perspectives

Our future work will focus on addressing the calibration of the outlier scores by introducing dependent loss functions since a false negative in smart agriculture and C-ITS scenarios can induce some uncomfortable issues especially on crop production, farm efficiency and road safety.

We also propose to introduce automatic parameter calibration with the aim of improving the algorithm's chances of deployment for farm and road infrastructure operators. Another critical part will be process optimization so as to gain in complexity with some fine adjustments of the ensemble learning decision rules for efficiency improvement.

The configuration of local subspaces can also be improved to decrease the amount of time spent locating a test instance's nearest neighbours by replacing the kNN search strategy with a clustering technique. We also propose to develop

the ELSCP methodology into a tool that can be used for real-time detection and analysis of data streams.

In the analysis of speed signatures and driving profiles, it will be of interest to do a more detailed analysis of the data to identify factors causing or influencing the observed behaviour, especially on the spike points. The DENM messages can also be used to extract points where incidents were reported and then use these locations as POIs for traffic incident detection. By comparing the extracted speed signatures of road segments against the known incident points from the DENMs, it could now be possible to get a better understanding on the evolution of driving behaviour.

To solve the data insufficiency problem in C-ITS, the speed signatures generated in real naturalistic driving environment in this study can be used to generate synthetic data from the learned characteristics of movement patterns of the vehicles. The signatures can also be applied in Autonomous vehicles where the signatures can be used as inputs in order to check if the driving is following the expected naturalistic speed behaviour.

There are a number of pseudonym change strategies in C-ITS which affect how identifiers are changed. The data used in this study applied a round robin strategy. It would be interesting to evaluate the data linking methodology developed in this thesis on data from other change strategies as a way of validating its efficacy.

Bibliography

- [1] W Sarni, J Mariani, and J Kaji. From dirt to data: The second green revolution and the internet of things. *Deloitte Review*, 18:4–19, January 2016.
- [2] Meng Lu, O Türetken, Onat Ege Adali, Jacint Castells, Robbin Blokpoel, and PWPJ Grefen. C-its (cooperative intelligent transport systems) deployment in europe: challenges and key findings. In *25th ITS World Congress*, pages EU–TP1076, 2018.
- [3] Shoaib Kamran and Olivier Haas. A multilevel traffic incidents detection approach: Identifying traffic patterns and vehicle behaviours using real-time gps data. In *2007 IEEE Intelligent Vehicles Symposium*, pages 912–917. IEEE, 2007.
- [4] Ibrahim Hassan Hashim. Analysis of speed characteristics for rural two-lane roads: A field study from minoufiya governorate, egypt. *Ain Shams Engineering Journal*, 2(1):43–52, 2011.
- [5] Cindie Andrieu, Guillaume Saint Pierre, and Xavier Bressaud. Estimation of space-speed profiles: A functional approach using smoothing splines. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 982–987. IEEE, 2013.
- [6] Juliet Chebet Moso, Ramzi Boutahala, Brice Leblanc, Hacène Fouchal, Cyril de Runz, Stephane Cormier, and John Wandeto. Anomaly detection on roads using c-its messages. In *International Workshop on Communication Technologies for Vehicles*, pages 25–38. Springer, November 2020.
- [7] Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John M Wandeto. Abnormal behavior detection in farming stream data. In *International Conference on Smart and Sustainable Agriculture*, volume 1470, pages 44–56. Springer. Cham, June 2021. doi: 10.1007/978-3-030-88259-4_4.

-
- [8] Juliet Chebet Moso, Stéphane Cormier, Cyril de Runz, Hacène Fouchal, and John Mwangi Wandeto. Anomaly detection on data streams for smart agriculture. *Agriculture*, 11(11), 2021. ISSN 2077-0472. doi: 10.3390/agriculture11111083. URL <https://www.mdpi.com/2077-0472/11/11/1083>.
- [9] Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Mwangi Wandeto. Streaming-based anomaly detection in its messages. submitted, 2021.
- [10] Juliet Chebet Moso, Stéphane Cormier, Hacene Fouchal, Cyril de Runz, John M Wandeto, and Hasnaâ Aniss. Road speed signatures from c-its messages. In *ICC 2021-IEEE International Conference on Communications*, pages 1–6. IEEE, June 2021.
- [11] Juliet Chebet Moso, Stéphane Cormier, Hacène Fouchal, Cyril de Runz, and John Wandeto. Trajectory user linking in c-its data analysis. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, December 2020.
- [12] Hasnaâ Aniss. Overview of an its project: Scoop@f. In *International Workshop on Communication Technologies for Vehicles*, pages 131–135. Springer, 2016.
- [13] Mohammad S Allahyari, Christos A Damalas, and Mehdi Ebadattalab. Farmers’ technical knowledge about integrated pest management (ipm) in olive production. *Agriculture*, 7(12):101, 2017.
- [14] Mario Fagnoli, Mara Lombardi, and Daniele Puri. Applying hierarchical task analysis to depict human safety errors during pesticide use in vineyard cultivation. *Agriculture*, 9(7):158, 2019.
- [15] Yaguang Zhang, Andrew Balmos, James V Krogmeier, and Dennis Buckmaster. Working zone identification for specialized micro transportation systems using gps tracks. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1779–1784. IEEE, 2015.
- [16] Muhammad Awais Javed and Elyes Ben Hamida. On the interrelation of security, qos, and safety in cooperative its. *IEEE Transactions on Intelligent Transportation Systems*, 18(7):1943–1957, 2016.

-
- [17] EN 302 663 V1. 2.1. Intelligent transport systems (its); access layer specification for intelligent transport systems operating in the 5 ghz frequency band. *ETSI*, July 2013.
- [18] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
- [19] Fabrizio Angiulli and Fabio Fassetti. Detecting distance-based outliers in streams of data. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 811–820, 2007.
- [20] Md Shiblee Sadik and Le Gruenwald. Dbod-ds: Distance based outlier detection for data streams. In *International Conference on Database and Expert Systems Applications*, pages 122–136. Springer, 2010.
- [21] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1067–1075, 2017.
- [22] Alban Siffer. *New statistical methods for data mining, contributions to anomaly detection and unimodality testing*. PhD thesis, Rennes 1, 2019.
- [23] Chun-Hsu Ou, Yan-An Chen, Ting-Wei Huang, and Nen-Fu Huang. Design and implementation of anomaly condition detection in agricultural iot platform system. In *2020 International Conference on Information Networking (ICOIN)*, pages 184–189. IEEE, 2020.
- [24] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft. Deepanomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors*, 16(11):1904, 2016.
- [25] Jie Xu, Suri Guga, Guangzhi Rong, Dao Riao, Xingpeng Liu, Kaiwei Li, and Jiquan Zhang. Estimation of frost hazard for tea tree in zhejiang province based on machine learning. *Agriculture*, 11(7):607, 2021.
- [26] Mustafa Abdallah, Wo Jae Lee, Nithin Raghunathan, Charilaos Mousoulis, John W Sutherland, and Saurabh Bagchi. Anomaly detection through transfer learning in agriculture and manufacturing iot systems. *arXiv preprint arXiv:2102.05814*, 2021.

- [27] Florian Mouret, Mohanad Albughdadi, Sylvie Duthoit, Denis Kouamé, Guillaume Rieu, and Jean-Yves Tourneret. Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and sar time series. *Remote Sensing*, 13(5):956, 2021.
- [28] Yang Wang, Andrew Balmos, James Krogmeier, and Dennis Buckmaster. Data-driven agricultural machinery activity anomaly detection and classification. In *Proceedings of the 14th International Conference on Precision Agriculture*, 2018.
- [29] Mingyang Zhang, Tong Li, Yue Yu, Yong Li, Pan Hui, and Yu Zheng. Urban anomaly analytics: Description, detection and prediction. *IEEE Transactions on Big Data*, 2020.
- [30] Shashi Shekhar, Hui Xiong, and Xun Zhou, editors. *Encyclopedia of GIS*. Springer, 2017. ISBN 978-3-319-17884-4. doi: 10.1007/978-3-319-17885-1. URL <https://doi.org/10.1007/978-3-319-17885-1>.
- [31] Francesco Alesiani, Luis Moreira-Matias, and Mahsa Faizrahmemon. On learning from inaccurate and incomplete traffic flow data. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3698–3708, 2018.
- [32] Mario Munoz-Organero, Ramona Ruiz-Blaquez, and Luis Sánchez-Fernández. Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on gps traces while driving. *Computers, Environment and Urban Systems*, 68:1–8, 2018.
- [33] Soodeh Dadras, Homayoun Jamshidi, Sara Dadras, and Thomas Edward Pilutti. Novel stop sign detection algorithm based on vehicle speed profile. In *2019 American Control Conference (ACC)*, pages 3994–3999. IEEE, 2019.
- [34] Hang Qiu, Jinzhu Chen, Shubham Jain, Yurong Jiang, Matt McCartney, Gorkem Kar, Fan Bai, Donald K Grimm, Marco Gruteser, and Ramesh Govindan. Towards robust vehicular context sensing. *IEEE Transactions on Vehicular Technology*, 67(3):1909–1922, 2017.
- [35] Jing Wang, Chaoliang Wang, Xianfeng Song, and Venkatesh Raghavan. Automatic intersection and traffic rule detection by mining motor-vehicle gps trajectories. *Computers, Environment and Urban Systems*, 64:19–29, 2017.

- [36] Eleonora D'Andrea and Francesco Marcelloni. Detection of traffic congestion and incidents from gps trace analysis. *Expert Systems with Applications*, 73: 43–56, 2017.
- [37] Qiang Gao, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Xucheng Luo, and Fengli Zhang. Identifying human mobility via trajectory embeddings. In *IJCAI*, volume 17, pages 1689–1695, 2017.
- [38] Andras Varga. The omnet++ discrete event simulation system. *Proceedings of the European Simulation Multiconference, June 2001*, pages 319–324, 2001. URL <https://ci.nii.ac.jp/naid/20001036826/en/>.
- [39] Daniel Krajzewicz, Georg Hertkorn, C. Rössel, and Peter Wagner. Sumo (simulation of urban mobility) - an open-source traffic simulation. In A. Al-Akaidi, editor, *4th Middle East Symposium on Simulation and Modelling*, pages 183–187, 2002. URL <https://elib.dlr.de/6661/>. LIDO-Berichtsjahr=2004,.
- [40] Raphael Riebl, Christina Obermaier, and Hendrik-Jörn Günther. Artery: Large scale simulation environment for its applications. In *Recent Advances in Network Simulation*, pages 365–406. Springer, 2019.
- [41] Kiana Hajebi, Yasin Abbasi-Yadkori, Hossein Shahbazi, and Hong Zhang. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *Twenty-Second International Joint Conference on Artificial Intelligence*, July 2011.
- [42] Rafael MO Cruz, Robert Sabourin, and George DC Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018.
- [43] Shebuti Rayana and Leman Akoglu. Less is more: Building selective anomaly ensembles. *Acm transactions on knowledge discovery from data (tkdd)*, 10(4):1–33, 2016.
- [44] Meghna Raj, Shashank Gupta, Vinay Chamola, Anubhav Elhence, Tanya Garg, Mohammed Atiquzzaman, and Dusit Niyato. A survey on the role of internet of things for adopting and promoting agriculture 4.0. *Journal of Network and Computer Applications*, page 103107, 2021.

-
- [45] Muhammad Fahim and Alberto Sillitti. Anomaly detection, analysis and prediction techniques in iot environment: A systematic literature review. *IEEE Access*, 7:81664–81681, 2019.
- [46] Redhwan Al-amri, Raja Kumar Murugesan, Mustafa Man, Alaa Fareed Abdulateef, Mohammed A Al-Sharafi, and Ammar Ahmed Alkahtani. A review of machine learning and deep learning techniques for anomaly detection in iot data. *Applied Sciences*, 11(12):5320, 2021.
- [47] Akanksha Toshniwal, Kavi Mahesh, and R Jayashree. Overview of anomaly detection techniques in machine learning. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 808–815. IEEE, 2020.
- [48] Yue Zhao, Zain Nasrullah, Maciej K Hryniewicki, and Zheng Li. Lscp: Locally selective combination in parallel outlier ensembles. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 585–593. SIAM, 2019.
- [49] Omar Alghushairy, Raed Alsini, Terence Soule, and Xiaogang Ma. A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, 5(1):1, 2021.
- [50] Reza Ehsani. Increasing field efficiency of farm machinery using gps. *EDIS*, 2010(5), 2010.
- [51] ETSI EN. 302 637-2 v1. 4.1-intelligent transport systems (its); vehicular communications; basic set of applications; part 2: Specification of cooperative awareness basic service. *ETSI*, April, 2019.
- [52] Andreas Festag. Cooperative intelligent transport systems standards in europe. *IEEE communications magazine*, 52(12):166–172, 2014.
- [53] ETSI TR. 103 415 v1.1.1 -intelligent transport systems (its); security; pre-standardization study on pseudonym change management. *ETSI*, April, 2018.
- [54] Verónica Saiz-Rubio and Francisco Rovira-Más. From smart farming towards agriculture 5.0: A review on crop data management. *Agronomy*, 10(2):207, 2020.

- [55] Miad Faezipour, Mehrdad Nourani, Adnan Saeed, and Sateesh Addepalli. Progress and challenges in intelligent vehicle area networks. *Communications of the ACM*, 55(2):90–100, 2012.
- [56] Muhammad Ayaz, Mohammad Ammad-Uddin, Zubair Sharif, Ali Mansour, and El-Hadi M Aggoune. Internet-of-things (iot)-based smart agriculture: Toward making the fields talk. *IEEE Access*, 7:129551–129583, 2019.
- [57] Hacène Fouchal, Emilien Bourdy, Geoffrey Wilhelm, and Marwane Ayaida. A framework for validation of cooperative intelligent transport systems. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2016.
- [58] Xudong Wang, Antoine Fagette, Pascal Sartelet, and Lijun Sun. A probabilistic tensor factorization approach to detect anomalies in spatiotemporal traffic activities. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1658–1663. IEEE, 2019.
- [59] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000, 2019.
- [60] Charu C Aggarwal. Outlier analysis. In *Data mining*, pages 237–263. Springer, 2015.
- [61] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012.
- [62] Andreas Kind, Marc Ph Stoecklin, and Xenofontas Dimitropoulos. Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management*, 6(2):110–121, 2009.
- [63] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
- [64] Mia Hubert, Michiel Debruyne, and Peter J Rousseeuw. Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3):e1421, 2018.

- [65] Peter J Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236, 2018.
- [66] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.
- [67] Wo-Ruo Chen, Yong-Huan Yun, Ming Wen, Hong-Mei Lu, Zhi-Min Zhang, and Yi-Zeng Liang. Representative subset selection and outlier detection via isolation forest. *Analytical Methods*, 8(39):7225–7231, 2016.
- [68] Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In *International conference on machine learning*, pages 2712–2721. PMLR, 2016.
- [69] Tomáš Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.
- [70] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.
- [71] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.
- [72] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004.
- [73] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123. IEEE, 2020. doi: 10.1109/ICDM50108.2020.00001.
- [74] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [75] Bouchra Lamrini, Augustin Gjini, Simon Daudin, Pascal Pratmarty, François Armando, and Louise Travé-Massuyès. Anomaly detection using

- similarity-based one-class svm for network traffic characterization. In *DX@ Safeprocess*, 2018.
- [76] Pedro Henriques dos Santos Teixeira and Ruy Luiz Milidiú. Data stream anomaly detection through principal subspace tracking. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1609–1616, 2010.
- [77] James Pickands III et al. Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131, 1975.
- [78] Mingming Zhang, Chao Chen, Tianyu Wo, Tao Xie, Md Zakirul Alam Bhuiyan, and Xuelian Lin. Safedrive: online driving anomaly detection from large-scale vehicle data. *IEEE Transactions on Industrial Informatics*, 13(4):2087–2096, 2017.
- [79] Simon Blackmore. The interpretation of trends from multiple yield maps. *Computers and electronics in agriculture*, 26(1):37–51, 2000.
- [80] G Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.
- [81] Simon Blackmore, Richard J Godwin, and Spyros Fountas. The analysis of spatial and temporal trends in yield map data over six years. *Biosystems engineering*, 84(4):455–466, 2003.
- [82] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1010–1018, 2011.
- [83] Linsey Xiaolin Pang, Sanjay Chawla, Wei Liu, and Yu Zheng. On detection of emerging anomalous traffic patterns using gps data. *Data & Knowledge Engineering*, 87:357–373, 2013.
- [84] Xiangjie Kong, Ximeng Song, Feng Xia, Haochen Guo, Jinzhong Wang, and Amr Tolba. Lotad: Long-term traffic anomaly detection based on crowd-sourced bus trajectory data. *World Wide Web*, 21(3):825–847, 2018.
- [85] Xiaolin Han, Tobias Grubenmann, Reynold Cheng, Sze Chun Wong, Xiaodong Li, and Wenya Sun. Traffic incident detection: A trajectory-based approach. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1866–1869. IEEE, 2020.

-
- [86] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):1–41, 2015.
- [87] Fabio Valdés and Ralf Hartmut Güting. A framework for efficient multi-attribute movement data analysis. *The VLDB Journal*, 28(4):427–449, 2019.
- [88] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–38, 2013.
- [89] A Nishad and Sajimon Abraham. Semtraclus: an algorithm for clustering and prioritizing semantic regions of spatio-temporal trajectories. *International Journal of Computers and Applications*, 43(8):841–850, 2021.
- [90] Yang Cao, Fei Xue, Yuanying Chi, Zhiming Ding, Limin Guo, Zhi Cai, and Hengliang Tang. Effective spatio-temporal semantic trajectory generation for similar pattern group identification. *International Journal of Machine Learning and Cybernetics*, 11(2):287–300, 2020.
- [91] Francisco Vicenzi, Lucas May Petry, Camila Leite da Silva, Luis Otavio Alvares, and Vania Bogorny. Exploring frequency-based approaches for efficient trajectory classification. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 624–631, 2020.
- [92] Maria Luisa Damiani, Fatima Hachem, Hamza Issa, Nathan Ranc, Paul Moorcroft, and Francesca Cagnacci. Cluster-based trajectory segmentation with local noise. *Data Mining and Knowledge Discovery*, 32(4):1017–1055, 2018.
- [93] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):1–32, 2013.
- [94] Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008.
- [95] Qingying Yu, Yonglong Luo, Chuanming Chen, and Shigang Chen. Trajectory similarity clustering based on multi-feature distance measurement. *Applied Intelligence*, 49(6):2315–2338, 2019.

- [96] BA Sabarish, R Karthi, and T Gireeshkumar. Clustering of trajectory data using hierarchical approaches. In *Computational Vision and Bio Inspired Computing*, pages 215–226. Springer, 2018.
- [97] Carlos Andres Ferrero, Luis Otavio Alvares, Willian Zalewski, and Vania Bogorny. Movelets: Exploring relevant subtrajectories for robust trajectory classification. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 849–856, 2018.
- [98] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. In *Proceedings 18th international conference on data engineering*, pages 673–684. IEEE, 2002.
- [99] Lei Chen, M Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502, 2005.
- [100] Hye-Young Kang, Joon-Seok Kim, and Ki-Joune Li. Similarity measures for trajectory of moving objects in cellular space. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1325–1330, 2009.
- [101] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S Tseng. Mining user similarity from semantic trajectories. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 19–26, 2010.
- [102] Andre Salvaro Furtado, Despina Kopanaki, Luis Otavio Alvares, and Vania Bogorny. Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, 20(2):280–298, 2016.
- [103] Andre L Lehmann, Luis Otavio Alvares, and Vania Bogorny. Smsm: a similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science*, 33(9):1847–1872, 2019.
- [104] Rune Elvik. Speed limits, enforcement, and health consequences. *Annual review of public health*, 33:225–238, 2012.
- [105] Heloisa M Barbosa, Miles R Tight, and Anthony D May. A model of speed profiles for traffic calmed roads. *Transportation Research Part A: Policy and Practice*, 34(2):103–123, 2000.

-
- [106] Aliaksei Laureshyn, Kalle Åström, and Karin Brundell-Freij. From speed profile data to analysis of behaviour: classification by pattern recognition techniques. *IATSS research*, 33(2):88–98, 2009.
- [107] Yann Méneroux, Arnaud Le Guilcher, Guillaume Saint Pierre, M Ghasemi Hamed, Sébastien Mustière, and Olivier Orfila. Traffic signal detection from in-vehicle gps speed profiles using functional data analysis and machine learning. *International Journal of Data Science and Analytics*, pages 1–19, 2019.
- [108] Brice Leblanc, Emilien Bourdy, Hacène Fouchal, Cyril de Runz, and Secil Ercan. Unsupervised driving profile detection using cooperative vehicles’ data. In *International Workshop on Communication Technologies for Vehicles*, pages 27–37. Springer, 2019.
- [109] Brice Leblanc, Hacene Fouchal, and Cyril de Runz. Driver profile detection using points of interest neighbourhood. In *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pages 1–4. IEEE, 2019.
- [110] Omar Chakroun and Soumaya Cherkaoui. Studying the impact of dsrc penetration rate on lane changing advisory application. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.
- [111] Fan Zhou, Qiang Gao, Goce Trajcevski, Kunpeng Zhang, Ting Zhong, and Fengli Zhang. Trajectory-user linking via variational autoencoder. In *IJCAI*, pages 3212–3218, 2018.
- [112] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. In *The World Wide Web Conference*, pages 459–469, 2019.
- [113] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- [114] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer, 2013.

-
- [115] Guilherme O Campos, Arthur Zimek, Jörg Sander, Ricardo JGB Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30(4):891–927, 2016.
- [116] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [117] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
- [118] Markus Ullmann, Thomas Strubbe, and Christian Wieschebrink. Technical limitations, and privacy shortcomings of the vehicle-to-vehicle communication. In *Fifth International Conference on Advances in Vehicular Systems*, 2016.
- [119] ETSI TS. 103 097 v1. 4.1-intelligent transport systems (its); security; security header and certificate formats. *ETSI, October*, 2020.
- [120] Albert Bifet and Richard Kirkby. Data stream mining a practical approach. 2009.
- [121] Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. Fast anomaly detection for streaming data. In *Twenty-Second International Joint Conference on Artificial Intelligence*, pages 1511–1516, 2011.
- [122] Mohamed Medhat Gaber. Advances in data stream mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):79–85, 2012.
- [123] Ece Calikus, Sławomir Nowaczyk, Anita Sant’Anna, and Onur Dikmen. No free lunch but a cheaper supper: A general framework for streaming anomaly detection. *Expert Systems with Applications*, 155:113453, 2020.
- [124] Alceu S Britto Jr, Robert Sabourin, and Luiz ES Oliveira. Dynamic selection of classifiers—a comprehensive review. *Pattern recognition*, 47(11):3665–3680, 2014.
- [125] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

- [126] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. Decision combination in multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence*, 16(1):66–75, 1994.
- [127] Kevin Woods, W. Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):405–410, 1997.
- [128] Shebuti Rayana and Leman Akoglu. An ensemble approach for event detection and characterization in dynamic graphs. In *ACM SIGKDD ODD Workshop*, 2014.
- [129] Arthur Zimek, Ricardo JGB Campello, and Jörg Sander. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter*, 15(1):11–22, 2014.
- [130] Charu C Aggarwal and Saket Sathe. Theoretical foundations and algorithms for outlier ensembles. *Acm sigkdd explorations newsletter*, 17(1):24–47, 2015.
- [131] Mohamad Dolatshah, Ali Hadian, and Behrouz Minaei-Bidgoli. Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces. *arXiv preprint arXiv:1511.00628*, 2015.
- [132] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*, chapter 4. Morgan Kaufmann, San Francisco, 2 edition, 2005.
- [133] Neeraj Kumar, Li Zhang, and Shree Nayar. What is a good nearest neighbors algorithm for finding similar patches in images? In *European conference on computer vision*, pages 364–378. Springer, 2008.
- [134] Yue Zhao and Maciej K Hryniewicki. Dcso: dynamic combination of detector scores for outlier ensembles. *arXiv preprint arXiv:1911.10418*, 2019.
- [135] RB Nelsen. Kendall tau metric. *Encyclopaedia of mathematics*, 3:226–227, 2001.
- [136] MG Kendall. Rank correlation methods 4th edition charles griffin. *High Wycombe, Bucks*, 1970.
- [137] Shravan Kumar Koninti. Av janatahack: Machine learning in agriculture, 2020. URL <https://www.kaggle.com/shravankoninti/av-janatahack-machine-learning-in-agriculture>.

- [138] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(1): 1–67, 2011.
- [139] Zhang Yaguang and Krogmeier James. Combine kart truck gps data archive, May 2020. URL <https://purr.purdue.edu/publications/3083/2>.
- [140] ASABE Standards. Agricultural machinery management data, asae d497.7(r2015). *American Society of Agricultural and Biological Engineers (ASABE)*, March 2011.
- [141] Shamilah Ahmad Mokhtor, Darius El Pebrian, and Nor Azi Asmindia Johari. Actual field speed of rice combine harvester and its influence on grain loss in malaysian paddy field. *Journal of the Saudi Society of Agricultural Sciences*, 19(6):422–425, 2020.
- [142] Zhenhua Zhang, Qing He, Hanghang Tong, Jizhan Gou, and Xiaoling Li. Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network. *Transportation Research Part C: Emerging Technologies*, 71:284–302, 2016.
- [143] Brice Leblanc, Hacene Fouchal, and Cyril De Runz. Obstacle detection based on cooperative-intelligent transport system data. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2020.
- [144] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.
- [145] Selim F Yilmaz and Suleyman S Kozat. Pysad: A streaming anomaly detection framework in python. *arXiv preprint arXiv:2009.02572*, 2020.
- [146] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*, 2019.
- [147] Arash Moradkhani Roshandeh, Mahmood Mahmoodi Nesheli, and Othman Che Puan. Evaluation of traffic characteristics: a case study. *International Journal of Recent Trends in Engineering*, 1(6):62–68, 2009.
- [148] Craig Norman Kloeden, Giulio Ponte, and Jack McLean. *Travelling speed and risk of crash involvement on rural roads*. Australian Transport Safety Bureau, 2001.

- [149] Ankit Kumar Yadav and Nagendra R Velaga. Investigating the effects of driving environment and driver characteristics on drivers' compliance with speed limits. *Traffic injury prevention*, 22(3):201–206, 2021.
- [150] Renata Torquato Steinbakk, Pål Ulleberg, Fridulv Sagberg, and Knut Inge Fostervold. Effects of roadwork characteristics and drivers' individual differences on speed preferences in a rural work zone. *Accident Analysis & Prevention*, 132:105263, 2019.
- [151] Yee Mun Lee, Siang Yew Chong, Karen Goonting, and Elizabeth Sheppard. The effect of speed limit credibility on drivers' speed choice. *Transportation research part F: traffic psychology and behaviour*, 45:43–53, 2017.
- [152] Mohd Khairul Alhapi Ibrahim, Hussain Hamid, Teik Hua Law, and Shaw Voon Wong. Use of continuous speed profiles to investigate motorcyclists' speed choice along exclusive motorcycle lane. In *IOP Conference Series: Materials Science and Engineering*, volume 512, page 012025. IOP Publishing, 2019.
- [153] Scoop project: connected road and vehicle. general presentation. Available at <http://www.scoop.developpement-durable.gouv.fr/en/general-presentation-a9.html>, 2018.
- [154] Simon Washington, Matthew Karlaftis, Fred Mannering, and Panagiotis Anastasopoulos. *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC, 2020.
- [155] Panagiotis Ch Anastasopoulos and Fred L Mannering. The effect of speed limits on drivers' choice of speed: a random parameters seemingly unrelated equations approach. *Analytic methods in accident research*, 10:1–11, 2016.
- [156] Tao Wu, Jianxin Qin, and Yiliang Wan. Tost: A topological semantic model for gps trajectories inside road networks. *ISPRS International Journal of Geo-Information*, 8(9):410, 2019.
- [157] Jonathan Petit, Florian Schaub, Michael Feiri, and Frank Kargl. Pseudonym schemes in vehicular networks: A survey. *IEEE communications surveys & tutorials*, 17(1):228–255, 2014.
- [158] Zhaojun Lu, Gang Qu, and Zhenglin Liu. A survey on recent advances in vehicular network security, trust, and privacy. *IEEE Transactions on*

Intelligent Transportation Systems, 20(2):760–776, February 2019. doi: 10.1109/TITS.2018.2818888.

Approches d'exploration des flux de données dans les systèmes de transport intelligents et l'agriculture de précision

Dans cette thèse, nous abordons le problème de l'analyse des données IoT en nous concentrant sur la détection des anomalies dans les flux de données et l'analyse des comportements. L'apprentissage non supervisé est intéressant pour les applications de la vraie vie, en particulier pour la détection des anomalies, car il y a beaucoup de données sans étiquettes dans ce scénario. Nous proposons une technique ELSCP (Enhanced Locally Selective Combination in Parallel outlier ensembles). Nous définissons une méthodologie non supervisée axée sur les données et nous l'appliquons à trois études de cas: la détection des dommages causés aux cultures dans un ensemble de données d'agriculture, l'application aux traces GPS des moissonneuses-batteuses et l'application aux messages issus des systèmes de transport intelligent coopératif (C-ITS). L'accent est mis sur l'identification des anomalies qui peuvent être liées à l'état ou à la santé des cultures pendant la récolte, celles qui ont un impact sur l'efficacité de la récolte et celles qui ont un impact sur la sécurité et le trafic routier. D'après nos résultats, il est possible de relier les anomalies extraites à l'état des cultures endommagées à la fin de la récolte. De même, nous avons été en mesure de détecter le comportement déviant de la moissonneuse-batteuse et d'identifier les anomalies sur les routes. Par conséquent, la détection des anomalies pourrait être intégrée dans le processus de décision des exploitants agricoles et routiers afin d'améliorer l'efficacité de la récolte, la santé des cultures, la sécurité routière et la fluidité du trafic.

Deuxièmement, nous avons considéré l'analyse des signatures de vitesse générées à partir des messages C-ITS dans le but de comprendre l'évolution du comportement de conduite dans un environnement de conduite naturelle. Nous avons montré qu'avec l'application de la segmentation et des statistiques agrégées, on est capable d'obtenir une meilleure compréhension du comportement général de conduite et de déduire des informations relatives à l'état de la route et à la situation du trafic routier. Enfin, nous avons examiné le problème de la liaison de trajectoires et l'avons appliqué aux messages C-ITS. Suite à notre analyse, il est possible de relier les trajectoires aux utilisateurs qui les ont générées si d'autres attributs discriminants et des connaissances de base sur la génération des messages sont pris en compte pendant l'analyse de similarité.

Mots-clés en français: Détection d'anomalies, Flux de données, Systèmes de transport intelligents, Agriculture intelligente, Détection d'incidents de circulation, Apprentissage non supervisé

Data Stream Mining Approaches in Intelligent Transportation Systems and Precision Agriculture

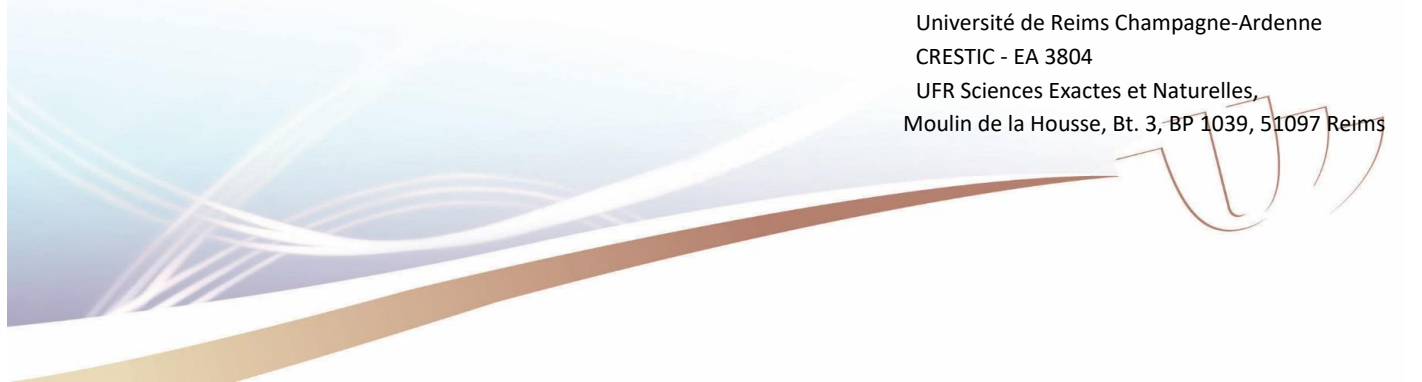
In this thesis, we address the problem of analysing IoT data with a focus on anomaly detection in data streams and behaviour analysis. Unsupervised learning is highly preferred for real-life applications, especially in anomaly detection since there is a lot of data without labels in this scenario. We propose an Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP) technique. We define an unsupervised data-driven methodology and apply it in three case studies; detection of crop damage in crop dataset, application to GPS logs of combine harvesters and application to Cooperative Intelligent Transport System (C-ITS) messages. The focus is the identification of anomalies that can be linked to crop state/health during harvest, those that have an impact on harvest efficiency and those impacting road safety and efficiency. Based on our results, it is possible to link anomalies extracted to damaged crop state at the end of harvest. Also, we were able to detect deviant behaviour of combine-harvester and to identify anomalies on the roads. Therefore, anomaly detection could be integrated in the decision process of farm and road operators to improve harvesting efficiency, crop health, road safety and traffic flow.

Secondly, we considered the analysis of speed signatures generated from C-ITS messages with the aim of understanding driving behaviour evolution under a naturalistic driving environment. We have shown that with the application of segmentation and aggregate statistics, one is able to get a better understanding of general driving behaviour and infer information that relates to the road condition and traffic situation. Finally, we considered the trajectory-linking problem and applied it to C-ITS messages. Based on our analysis, it is possible to link trajectories to the generating users if other distinguishing attributes and background knowledge on generation of the messages are considered during similarity analysis.

Mots-clés en anglais: Anomaly detection, Data streams, Intelligent Transportation Systems, Smart farming, Traffic incident detection, Unsupervised learning

Discipline : INFORMATIQUE – COMPUTER SCIENCE

Spécialité : Informatique – Computer Science



Université de Reims Champagne-Ardenne
CRESTIC - EA 3804
UFR Sciences Exactes et Naturelles,
Moulin de la Housse, Bt. 3, BP 1039, 51097 Reims