+

Diagnostic d'Infrastructures et Dynamique du Véhicule pour les Motos et les Autos

Livrable L6

Comparaison d'une approche de détection d'événements par dépassement de seuils dynamiques avec une approche basée sur des algorithmes d'apprentissage par réseaux de neurones

N° Livrable	**L6**	**N° workpackage**	**WP6**
Statut		Version provisoire	
Date		13/04/2022	
Responsable du document		**Thierry Serre, Université Gustave Eiffel** thierry.serre@univ-eiffel.fr	
Auteurs principaux		**Jérémie Fourmann, Liberty Rider** **Thierry Serre, Université Gustave Eiffel**	
Contributeur(s)		Sélim Cheaibi Christophe Perrin Claire Naude Jean-Yves Fournier	
Validation		Thierry Serre	
Enregistrement		DYMOA+_L6.doc	

Résumé

Le nombre croissant d'usagers de deux-roues motorisés (2RM) dans la mobilité urbaine et l'accidentologie représente un enjeu majeur en termes de sécurité. Avec plus de 23% du total des tués et 31% du total des blessés, pourcentages importants, alors même que l'on estime à 1.5% la part de ces véhicules dans le trafic total (ONISR, 2006). La compréhension des interactions des 2RM avec d'une part les autres usagers et d'autre part l'infrastructure routière est un des enjeux majeurs de la sécurité routière. Le projet DYMOA+ s'inscrit dans la continuité du projet DYMOA (achevé en 2017) et a pour objectif de produire de la connaissance manquante dans le domaine grâce à l'exploitation des résultats et des données expérimentales enregistrées dans le projet.

Le Work Package 6 (WP6) objet de ce rapport s'intéresse à ce que peuvent apporter les méthodes d'apprentissage supervisé" (machine learning) par rapport à des méthodes précédemment développées (projet DYMOA 2017) pour la classification de situations de danger dans lesquelles l'usager de 2RM doit réaliser une manœuvre d'urgence afin d'éviter un potentiel accident. En agrégeant ces situations, il est alors possible d'identifier des zones à risques pour l'usager du 2RM.

Cette étude a permis de montrer que les méthodes d'apprentissage supervisé sont adaptées à la détection d'une situation de danger dans laquelle l'usager doit effectuer une manœuvre d'urgence afin d'éviter l'accident. Des performances de l'ordre de 94% ont été obtenues en termes de classification sur les mesures de l'expérimentation du projet DYMOA. Par ailleurs l'étude a permis de montrer que l'interprétation des décisions des modèles permet de retrouver des similitudes avec la méthode de détection par seuils dynamiques développée précédemment dans le projet DYMOA. Cette méthode a également montré des limitations quant à l'optimisation de ses performances du fait de son interdépendance avec la qualité de la labellisation des données.

L'identification de ce type de situations permet de mettre en place une méthodologie d'identification des zones d'accumulation d'incidents lors de l'usage du 2RM. Cependant, des limitations sont apparues quant à l'étude des liens de causalité qui pourraient expliquer le caractère à risque de la zone. Chaque situation complexe nécessite le regard d'un expert humain et une étude détaillée de la scène afin de dresser le bilan des caractéristiques de cette zone à risque.

Les techniques d'apprentissage supervisé ainsi détaillées sont donc des outils complémentaires qui présentent de nombreux avantages et sont prometteuses pour le futur car elles pourraient permettre de mettre en lumière des zones potentiellement à risque qui pourront être analysées plus en profondeur avec l'expertise humaine.

Table des matières

1. Introduction

1.1 Présentation du Work Package 6

Le nombre croissant de 2RM dans la mobilité urbaine et l'accidentologie représente un enjeu majeur en termes de sécurité. En effet, le 2RM constitue un moyen de déplacement de plus en plus prisé mais reste encore à ce jour, malgré une baisse générale de l'accidentalité, un mode de transport particulièrement dangereux et ses utilisateurs des usagers très vulnérables. Le nombre de conducteurs de 2RM victimes d'accidents représente en France plus de 23% du total des tués (15% pour l'ensemble de l'Europe) et 31% du total des blessés, alors même que l'on estime à 1.5% la part de ces véhicules dans le trafic total, en termes de kilomètres parcourus (ONISR, 2006). Et quelles que soient les mesures prises ces dernières années, elles ne sont pas parvenues à faire décroître significativement ces taux. De plus, les 2RM sont la catégorie de véhicule qui a le moins bénéficié des importants progrès de sécurité routière constatés en France au cours de la dernière décennie. On déplore aussi un manque de connaissances générales sur les interactions des 2RM avec d'une part les autres usagers et d'autre part l'infrastructure routière.

Ce déficit de connaissance a été l'origine du projet DYMOA (achevé en 2017), dont les objectifs du projet étaient :

- Développer de nouvelles méthodes de diagnostic des infrastructures routières et de leur usage par des 2RM (Deux-Roues Motorisés) et des VL à l'aide d'EDR (Enregistreurs de Données de la Route), basées notamment sur l'analyse des incidents.

- Produire de la connaissance sur l'utilisation réelle d'un 2RM, en distinguant les interactions avec l'infrastructure, l'utilisation des capacités dynamiques des 2RM

- Mettre en œuvre une méthodologie de recueil (EDR de type smartphone, base de données) et d'exploitation de données (outils cartographiques) en conformité avec les droits des conducteurs concernés (protection des données à caractère personnel).

L'expérimentation a permis de recueillir plus de 6000 parcours pour une distance de plusieurs dizaines de milliers de kilomètres issues d'une flotte de 26 motocyclistes volontaires qui ont circulé principalement dans les 4 départements suivants : Eure, Seine-Maritime, Bouches-du-Rhône et Hérault. Les analyses ont apporté des informations originales sur la conduite des 2RM avec le recueil d'événements et d'incidents de conduites incluant les éléments de contexte obtenus à l'aide de vidéo embarquée.

Le projet DYMOA+ qui a débuté en 2019, a pour principal objectif de réaliser des exploitations complémentaires des résultats et données du projet DYMOA. Le Work Package 6 (WP6) s'intéresse à ce que peut apporter des méthodes en apprentissage supervisé (machine learning) par rapport à des méthodes précédemment développées dans le projet DYMOA pour la détection de zone à risques pour l'usager du deux-roues motorisés.

Depuis quelques années, de nombreux modèles basés sur la technique d'apprentissage supervisé et sur l'intelligence artificielle permettent d'analyser et de comprendre des situations complexes jusqu'alors réservées aux savoir-faire de quelques experts. Par exemple, en analysant les signaux d'un électrocardiogramme (ECG) depuis une montre connectée, il est possible de prévenir l'approche d'une crise cardiaque. En analysant la centrale inertielle de la montre, il est envisageable de détecter précisément le type d'activité que la personne est en train d'exécuter (marche, course, natation, etc) et détecter une situation anormale comme une noyade ou une chute. Pour la voiture autonome, de nombreux algorithmes ont pu être développés et sont capables de traiter des données vidéo et de nombreux signaux de capteurs embarqués afin de comprendre la situation, pouvant ainsi anticiper les interactions entre les véhicules et les alentours, et permettant de pouvoir choisir sa trajectoire de façon optimale sur le réseau routier.

L'idée principale de l'étude menée dans le WP6 est d'expérimenter ces techniques d'apprentissage supervisé sur les données de l'utilisation réelle d'un 2RM recueillies avec les EMMAPhones dans le projet DYMOA. En analysant les signaux enregistrés par l'EDR et en utilisant ces méthodes d'apprentissage supervisé, il serait envisageable de pouvoir détecter des incidents liés à l'usage du 2RM.

Le périmètre du projet est limité à l'étude des événements et des incidents pouvant apparaître dans l'usage du 2RM. Voici ci-dessous les définitions utilisées pour le reste de l'étude :

Événement : sollicitation forte mais très courte qui apparaît dans une situation de conduite tout à fait normale et où il n'y a pas a priori de caractère de danger. Un événement est principalement lié à l'infrastructure (passage sur un ralentisseur, dos d'âne, défaut de la route, etc).

Incident : forte sollicitation provoquée par l'action du motocycliste dans une manœuvre volontaire dite d'urgence (évitement ou freinage) afin d'éviter une situation de danger imminent qui aurait pu mener à une situation d'accident. Un incident n'est donc pas un accident.

Par ailleurs, en analysant de manière géographique les lieux et concentrations de ces événements/incidents, il peut être possible de mettre en évidence des zones de l'infrastructure routière sujettes à un risque plus élevé en termes d'accidentalité. La mise en évidence de ces zones à risque pourrait permettre dans le futur d'étudier plus en détail l'infrastructure pour comprendre l'usage du 2RM dans le but de prévenir de potentiels accidents.

Des applications industrielles pourraient voir le jour et permettraient notamment de fournir à la collectivité des outils cartographiant ces zones à risque en temps réel, et des outils d'alerte et de prévention pour l'usager de 2RM, permettant par exemple l'apparition d'alertes "zone à risque" dans les applications de navigation ou appareils GPS afin d'avertir avant l'entrée dans une zone où une vigilance particulière serait recommandée.

1.2 Contributeurs de l'étude

1.2.1 Le LMA de L'Université Gustave Eiffel

L'Université Gustave Eiffel est une université âgée de moins de deux ans. Fruit de la fusion de l'Université Paris-Est-Marne-la-Vallée (UPEM) et de l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR). Elle rassemble plus de 17 000 étudiants, 600 enseignants et enseignants-chercheurs et près de 550 chercheurs. Cette université multidisciplinaire est composée de 37 laboratoires répartis entre 5 départements de recherche : AME (Aménagement, Mobilité, Environnement), MAST (Matériaux et Structures), GERS (Géotechnique, Environnement, Risques naturels et Sciences de la Terre), COSYS (Composants et Systèmes), TS2 (Transport, Santé, Sécurité).

Le LMA est un laboratoire du département TS2, situé à Salon-de-Provence. L'activité scientifique porte sur l'étude des mécanismes générateurs d'accidents, des processus de dysfonctionnement du système de circulation et l'analyse de l'insécurité routière et cherche à proposer des aides à la conception (aménagement, véhicule) et à la formation (concepteurs, aménageurs, usagers). Il regroupe une équipe pluridisciplinaire de spécialistes en psychologie, mécanique et dynamique du véhicule, génie civil et urbanisme, en droit et sciences politiques.

1.2.2 Liberty Rider

La société Liberty Rider édite une application moto qui détecte les accidents et envoie automatiquement les secours en cas de chute du motard. L'application Liberty Rider, installée sur le smartphone du motard, détermine s'il a chuté ou non et si c'est le cas déclenche une alerte. Si le motard ne réagit pas, Liberty Rider prend la situation en main via son centre d'appel 24/7. Le centre d'appel va tenter d'appeler le motard 3 fois. Si celui-ci ne répond pas, Liberty Rider prévient les secours qui sont envoyés sur les lieux de l'accident. La procédure normée a permis de réduire considérablement les temps d'intervention.

L’innovation autour de la prévention et la sécurité routière est donc le moteur de Liberty Rider. Elle propose également un écosystème complet autour de l’usage du deux-roues motorisés comme par exemple sa navigation GPS spécialement conçue pour l’usage du 2RM qui alerte de l'approche de virage dangereux, mais aussi une plateforme qui recense les plus belles sorties et itinéraires autour de soi, et enfin un carnet d’entretien mécanique qui permet au motard d'être alerté à l’échéance de ses prochains entretiens.

L’identification des zones à risque pour l’usager de 2RM par des nouvelles méthodes en apprentissage supervisé est un sujet qui va prendre de plus en plus d'ampleur sur les années futures selon Liberty Rider. En effet, avec l'arrivée progressive de différentes innovations autour du 2RM et des véhicules connectés, on peut espérer qu’il sera un jour possible de prévenir l’usager en temps réel de l'arrivée dans une zone à risque afin qu’il puisse être tout particulièrement vigilant et qu’il évite l’accident.

1.3 Objectif et plan de l’étude

L'étude présentée dans ce rapport est composée de deux parties. La première partie de l’étude a pour objectif de comparer les approches à base d’apprentissage supervisé (machine learning) par rapport à l’approche de dépassement de seuils dynamiques (développée dans le projet DYMOA) pour la détection événements/incidents.

Les principales questions traitées dans cette partie sont :

- Est-ce que les techniques d’apprentissage sont appropriées pour la détection d’incident et d'évènement lié à l’usage de deux-roues motorisé ?

- Quelles différences ont ces méthodes par rapport aux méthodes dites classiques basées sur le dépassement de seuils ?

- Quelles sont les performances et les limites de l’apprentissage supervisé pour ce type de sujet ?

La seconde partie de l’étude s'intéresse à l'identification des zones potentiellement à risque lorsqu’il y a une accumulation d’incidents et d'évènements détectés par des méthodologies et sources différentes. Une fois que ces incidents et événements sont détectés, il est alors possible de localiser des zones à risque. Des techniques basées sur l'agrégation géographique et sur l’analyse statistique pourraient nous permettre d'identifier des zones à risques pour l’usager de 2RM. Cette partie analysera également les zones d'accumulation autour des accidents corporels de 2RM renseignés dans le fichier national des accidents corporels de la circulation dit « Fichier BAAC », disponible en libre accès sur la plateforme data.gouv.fr et administré par l’Observatoire national interministériel de la sécurité routière (ONISR).

Les principales questions auxquelles cette seconde partie vise à répondre sont :

- Est-il possible d’identifier des zones géographiques d’accumulation en agrégeant plusieurs sources de données ?

- Une fois la zone identifiée, est-il possible de comprendre les liens de causalité qui font de cette zone un lieu à forte sollicitations dynamiques ?

- Est-ce que les zones identifiées nous permettent de faire des liens entre accidentalité, événements et incidents ?

2. Partie 1 : Comparaison des méthodes d'apprentissage supervisé et des méthodes de dépassement de seuils dynamiques

L'objectif principal de cette partie est de comparer des méthodes d'apprentissage supervisé permettant de détecter une situation d'événement ou d'incident à partir d'enregistrements provenant de motos instrumentées. Les données utilisées proviennent de l'enregistrement dynamique de 26 2RM dans les départements de l'Eure, la Seine Maritime, des Bouches-du-Rhône et de l'Hérault, sur une période d'un an et demi (18 mois) pendant laquelle 6500 parcours ont été enregistrés représentant près de 80 000 km parcourus issus du projet DYMOA.

2.1. Présentation du dataset

Les données collectées par les EMMAPhones lors de la campagne d'expérimentation du projet DYMOA en 2017-2018 proviennent du capteur GPS et de la centrale inertielle du téléphone, placé sous la selle du 2RM. L'enregistreur est placé dans le repère lié à la moto, défini par la figure 1-A.

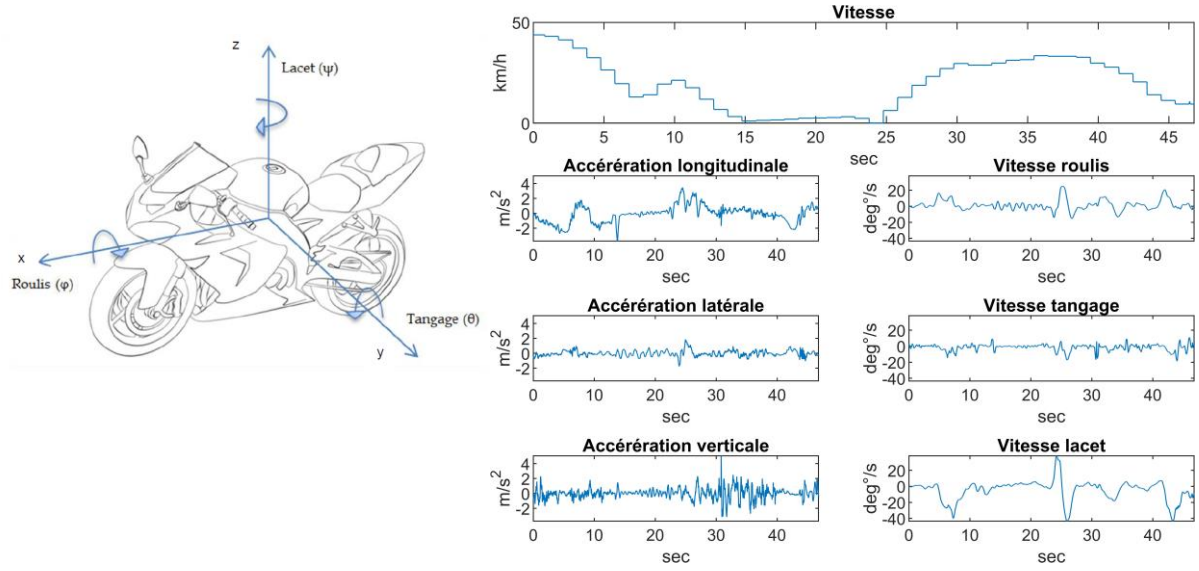

Figure 1 : A - Repère utilisé par l'EDR placé sous la selle de la moto ; B - Représentation des données primaires enregistrées par un EDR durant 45s (DYMOA)

Les données primaires collectées et présentées dans la figure 1-B sont :

- La vitesse (en km/h) fournie à une fréquence de 1Hz
- Les signaux de l'accéléromètre fournis à une fréquence de 100Hz avec les composantes suivantes :
  - La composante longitudinale de l'accélération suivant X en m/s² (positive en accélération, négative en freinage)
  - La composante transversale de l'accélération suivant Y en m/s² (positive en virage à gauche et négative en virage à droite)
  - La composante verticale de l'accélération suivant Z en m/s²
- Les signaux du gyroscope fournis à une fréquence de 100Hz
  - la vitesse de roulis en °/s
  - la vitesse de tangage en °/s
  - la vitesse de lacet en °/s

L’expérimentation du projet DYMOA sur les 26 volontaires conducteurs de 2RM a permis d'enregistrer 3800 déclenchements. Ces déclenchements ont fait l’objet d’un travail de labellisation afin de les regrouper dans 3 catégories distinctes.

Catégorie de type “Roulage” : déclenchement sur critères géographiques et correspondant à une phase de roulage sans dépassement des seuils

Catégorie de type "Événement" : déclenchement par dépassement des seuils lié à la sollicitation mécanique de courte durée lors d’une interaction avec une infrastructure (passage d’un ralentisseur, dos d’âne ou défaut de la route). La situation ne présente pas de caractère de danger a priori.

Catégorie de type “Incident” : déclenchement par dépassement de seuils, lié à des sollicitations dynamiques fortes lors d’une manœuvre d’urgence (freinage ou d'évitement) potentiellement liée à d’un danger imminent pour éviter une situation d’accident.

Le tableau 1 permet de résumer les données disponibles pour l’étude. On notera que la majorité des freinages forts sont labellisés en tant que incidents et que la plupart des fortes secousses sont labellisées en tant qu'événements.

Tableau 1 : Résumé des données labellisées par catégorie et par cause

Label	Cause	Compte	Pourcentage
Roulage	Critère géographique	551	100.00%
Evénement	Forte secousse	1801	99.67%
	Freinage fort	1	0.06%
	Forte vitesse de roulis	5	0.28%
Incident	Freinage fort	1078	80.33%
	Forte secousse	239	17.81%
	Forte vitesse de roulis	24	1.79%
	Forte vitesse de lacet	1	0.07%

L’étude porte sur la prédiction par des méthodes d’apprentissage supervisé des catégories “Roulage”, "Événement" et “Incident”. Les causes de ces déclenchements listés dans le tableau 1 apporte une information supplémentaire mais ne feront pas l’objet de prédiction par des méthodes d’apprentissage supervisé. La volumétrie des données nous permet d’envisager d’utiliser des méthodes d'apprentissage supervisé (type machine learning). Une attention particulière sera apportée sur la sélection des types de modèle afin d’éviter les problèmes de surapprentissage et de non généralisation des performances liées à un volume de données trop faible. Une partie de ces données (environ 70%) servira à l’apprentissage du modèle tandis qu’une autre partie (30%) sera utilisée pour l'évaluation des performances. Le choix de la catégorisation d’une situation en “Roulage”, "Événement" et “Incident” est réalisé par un humain suivant une définition donnée. Il n’est pas toujours facile d'associer une situation à sa catégorie, et par conséquent les données utilisées peuvent contenir des erreurs. Sous un certain niveau d’acceptabilité ses erreurs n’ont que peu de conséquence sur les performances et l’apprentissage de modèle de type machine learning.

Cependant, si la définition des catégories évolue, le modèle entraîné ne sera plus valable, il faudra ré-entraîner et évaluer les performances d’un nouveau modèle. La méthodologie d’apprentissage supervisé utilisée est présentée dans la section ci-dessous.

2.2. Méthode d'apprentissage supervisé pour la détection d'événements/incidents en deux-roues motorisés

2.2.1 Principe général de l'apprentissage supervisé

Les algorithmes d'apprentissage supervisé sont des outils du machine learning. Ils sont utilisés sur des données sur lesquelles on cherche à extraire des connaissances qu'on utilise ensuite pour définir des règles de classement.

La figure 2 représente les différentes étapes à suivre pour le développement d'un algorithme d'apprentissage supervisé. On distingue une phase dite d'apprentissage supervisé, et une phase prédictive qui utilise le modèle entraîné pour faire une prédiction avec des données d'entrées et permettant de trouver la catégorie associée.

Cette méthode est très largement utilisée dans de nombreux domaines (finance, marketing, médecine, automobile) et cela depuis une dizaine d'années. L'essentiel de ces techniques reposent sur la qualité et le volume de données préalablement labellisées par des humains et experts du domaine.

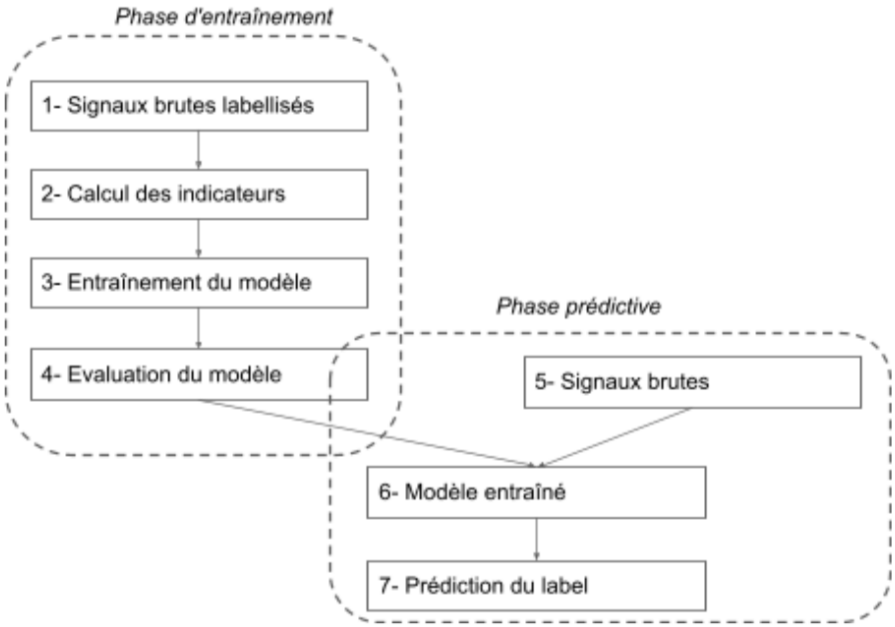

Figure 2 : A - Méthodologie d'apprentissage supervisé utilisé dans l'étude

2.2.2 Présentation des modèles utilisés dans l'étude

Les modèles utilisés dans cette étude doivent répondre aux contraintes suivantes :

- Pouvoir apprendre sur un volume de données compatible avec celui de l'étude,
- Être un modèle stable et facilement généralisable,
- Être un modèle dont la prise de décision est interprétable et compréhensible.

La première contrainte porte sur le choix d'un modèle par rapport au volume de données disponible pour l'apprentissage. En effet certains types de modèle demande des volumes de données gigantesques (plusieurs millions ou giga d'observations) afin d'espérer avoir des résultats convenables. Il est nécessaire que le choix de nos modèles puisse être cohérent avec la

volumétrie des observations précédemment récoltées dans la campagne d'expérimentation du projet DYMOA. Ensuite la deuxième contrainte porte sur la capacité de stabilité et de généralisation du modèle vis à vis d'une nouvelle observation qui n'aurait pas été enregistrée et présentée lors de la phase d'entraînement. Cela permet de s'assurer que l'on a pu faire apprendre le modèle à traiter le problème dans son ensemble et non juste par cœur à reconnaître des données d'entraînement. La troisième contrainte vise à pouvoir interpréter et comprendre la prédiction du modèle. Cela permet d'éviter l'effet "boîte noire" qui engendre des résultats difficilement compréhensibles et non améliorables. Un modèle interprétable permet également de réévaluer des indicateurs pertinents pour les approches de détection dite à dépassement de seuils dynamiques qui fera l'objet d'un développement plus approfondi dans l'étude WP5.

Au vu de ces critères les modèles sélectionnés pour l'étude sont le KNN, le SVM et l'Arbre de décision.

Modèle KNN

Le modèle des KNN [1] (K-nearest neighbors) est un modèle à 1 paramètre K. Au moment du test pour chaque donnée, il va regarder le label de ses K plus proches voisins dans l'espace et prendre le même label que la majorité de ses voisins. Pour ce modèle, il est recommandé d'utiliser un K impair.

Les principaux avantages sont :

- L'algorithme est simple et facile à implémenter.
- Il n'est pas nécessaire de créer un modèle, de régler plusieurs paramètres ou de formuler des hypothèses supplémentaires.
- L'algorithme est polyvalent. Il peut être utilisé pour la classification ou la régression.

Les principaux inconvénients sont :

- L'algorithme devient beaucoup plus lent à mesure que le volume de donnée d'apprentissage augmente

Modèle SVM

Le modèle d'apprentissage utilisant la technique des SVM [2] (Support Vector Machine) va définir une marge dans l'espace permettant de maximiser la distance entre la frontière des clusters.

Les principaux avantages sont :

- Grande précision de prédiction avec de petits volumes de données d'apprentissage
- Existence d'un paramètre permettant de régler la marge inter-cluster

Les principaux inconvénients sont :

- Temps d'entraînement long sur de gros volumes de données
- Manque de robustesse aux bruits et aux points aberrants

Modèle arbre de décision

Un arbre de décision [3] fonctionne en appliquant de manière itérative des règles logiques permettant de séparer de façon optimale les données lors de l'entraînement. Le grand gain des arbres de décision par rapport à une autre méthode de l'apprentissage supervisée est qu'il est très simple de l'interpréter pour analyser les résultats. On peut choisir la profondeur de l'arbre. Comme il n'utilise pas de distance, c'est le seul modèle étudié ici pour lequel il n'est pas nécessaire de normaliser les données. Plus un nœud se trouve haut dans l'arbre et plus il va discriminer un grand nombre de données.

Les principaux avantages sont :

- Facilité d'entraînement

- Grande interprétabilité

Les principaux inconvénients sont :

- Tendance au sur-apprentissage et à la non généralisation
- Pas adapté au faible volume de donnée

2.3. Résultats

2.3.1. Sélection des variables

L'ensemble de ces modèles vont devoir prédire la catégorie de l'observation représentée par un vecteur d'entrée de dimension N. Chaque dimension est déterminée par un indicateur appelé couramment "Variable" ou "Feature". Chaque indicateur est issu d'un traitement des données brutes puis ajouté dans un vecteur d'entrée servant à la prédiction (Figure 3).

Figure 3 : Passage des signaux brutes au vecteur d'entrée utilisé lors de l'apprentissage

Les principaux indicateurs utilisés dans l'étude sont des descripteurs statistiques tels que le maximum, le minimum, la valeur moyenne, la médiane, appliqués sur les signaux :

- Accélération linéaire (X, Y, Z)
- Vitesse de rotation (X, Y, Z)
- Accélération de rotation
- Dérivé de l'accélération (Jerk)
- Intervalle d'analyse autour du déclenchement

Exemple : l'indicateur "MaximumNormAccGyr22" correspond au maximum de la norme de l'accélération de rotation sur l'intervalle -2s avant et +2 secondes après le déclenchement.

Une fois les vecteurs d'entrées calculés pour chaque observation, il est possible de visualiser les données de l'étude en les projetant dans un espace à 2 dimensions. On peut ainsi analyser la dispersion et le regroupement des données sous la forme de cluster de points en fonction de leur catégorie "roulage", "événement" et "incident". La figure 4 représente la projection des données sur 2 dimensions.

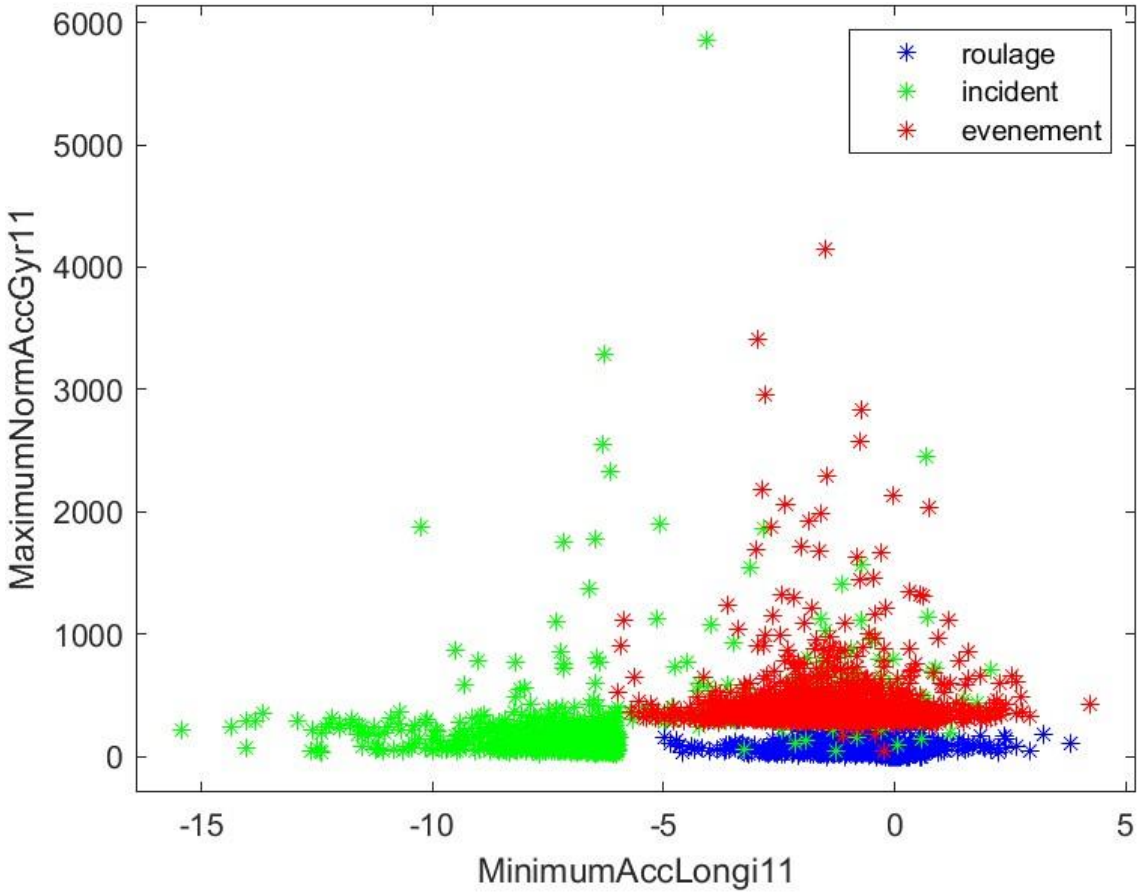

Figure 4 : Projection des observations sur 2 indicateurs (le minimum de l'accélération longitudinale et le maximum de la norme de vitesse de rotation)

Nous observons que les données semblent former des clusters identifiables mais se recouvrent à leurs frontières. C'est là où les techniques d'apprentissage supervisées vont pouvoir identifier ces clusters de façon performante dans des espaces à plusieurs dimensions.

Une fois les données sélectionnées et les vecteurs d'entrées assemblés, il faut réduire le nombre de variables pris en compte avant de pouvoir utiliser les modèles d'apprentissage supervisé. De nombreuses approches déjà programmées sur l'outil Matlab ont été utilisées. Pour se faire, il faut se rappeler qu'il est aisé de remonter un arbre de décision et de retrouver les nœuds et leurs niveaux dans l'arbre. Ainsi, l'idée de cette sélection de variables est très simple : on génère un arbre de décision à partir de deux tiers des données prises aléatoirement dans notre tableau (assemblé en fonction des données et des variables). On regarde quelles variables ont été utilisées et à quel niveau dans l'arbre. On répète le processus 100 fois pour construire 2 vecteurs nous renseignant sur l'importance des indicateurs dans la prédiction du label. On retrouve dans le graphe de la figure 5 un exemple de l'utilisation de nos vecteurs avec en orange les niveaux dans l'arbre à chaque temps et en bleu, le nombre d'utilisations sur les 100 répétitions

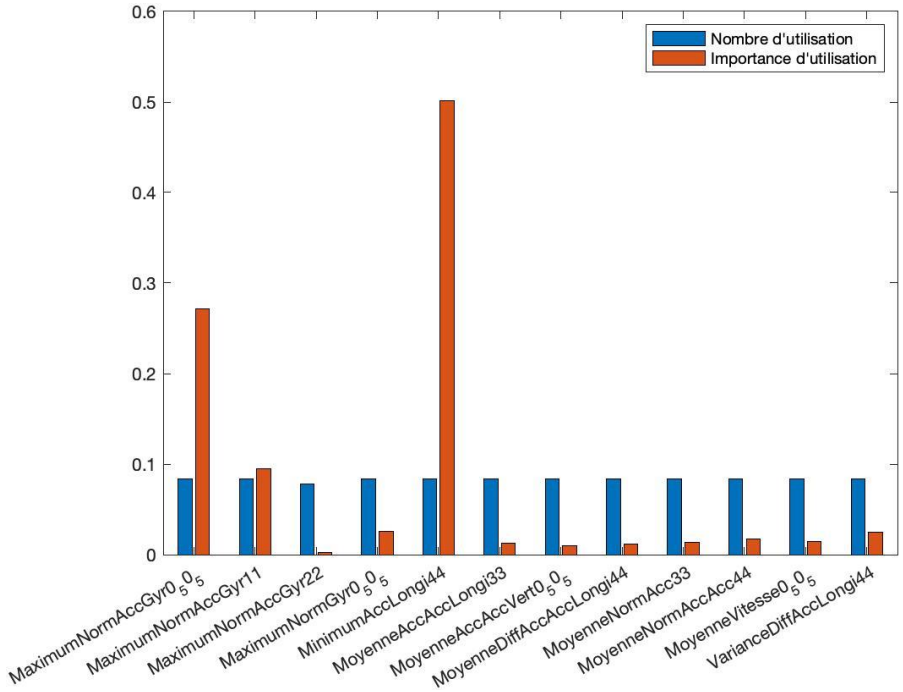

Figure 5 : Variables identifiées comme les plus discriminantes pour la classification

2.3.2. Comparaison des performances des modèles

L'estimation de la performance des modèles d'apprentissage supervisé se déroule via la méthodologie de la validation croisée. Dans notre cas, il s'agit de récupérer deux tiers de nos données, s'en servir pour l'entraînement d'un modèle d'apprentissage supervisé et d'utiliser le tiers restant pour tester l'algorithme et observer sa précision. On réitère l'opération à 100 reprises de façon aléatoire afin que la fonction de sélection des données ne biaise pas notre estimation de performance.

A partir des 12 variables sélectionnées, on entraîne plusieurs modèles en parallèle avec plus ou moins de variables d'entrées et on estime la performance de chaque modèle via la méthode de la validation croisée. Les résultats de ces tests sont rassemblés dans le Tableau 2. Les modèles de type KNN n'ont pas donné de résultats performants, ils n'ont donc pas été retenus dans la suite de l'étude. Les arbres de décision et SVM ont quant à eux obtenus les meilleurs résultats.

Tableau 2 : Résumé des performances des différents modèles entraînés

Sélection	Minimum	Maximum	Moyenne	Variance
SVM_11	92,35 %	95,52 %	93,59 %	0,00003
SVM_10	91,53 %	95,11 %	93,56 %	0,00004
SVM_4	89,01 %	94,95 %	93,50 %	0,00006
SVM_12	92,02 %	95,44 %	93,46 %	0,00004
SVM_3	91,21 %	94,87 %	93,44 %	0,00003
SVM_5	90,64 %	94,79 %	93,31 %	0,00004
SVM_9	91,78 %	94,79 %	93,30 %	0,00004
SVM_8	92,02 %	95,11 %	93,29 %	0,00004
SVM_6	91,69 %	94,95 %	93,27 %	0,00005
SVM_2	91,69 %	94,79 %	93,16 %	0,00003
SVM_7	91,37 %	94,71 %	93,16 %	0,00004
arbre_12	90,88 %	94,30 %	92,86 %	0,00004
arbre_10	90,88 %	94,22 %	92,81 %	0,00004
arbre_11	90,88 %	94,38 %	92,80 %	0,00004
arbre_9	90,15 %	94,46 %	92,64 %	0,00005
Originel	90,15 %	93,40 %	91,83 %	0,00005

On peut lire dans ce tableau que les résultats pour 1 à 12 variables sont plutôt équivalents, que les SVM ont de meilleurs résultats que les autres méthodes et que la précision moyenne de nos algorithmes est proche de 94% avec les arbres de décision ce qui est une précision remarquable. Par ailleurs l'écart entre la performance minimale et maximale de chaque modèle entraîné est relativement faible (inférieur à 5%) ce qui caractérise la stabilité des modèles entraînés lors de l'étude. Une sélection de variables est cependant pertinente (Cf. L5) avec nos outils puisque nous gagnons environ 2% de précision entre la meilleure sélection de variables (93.6%) et les quatre indicateurs utilisés originellement dans le projet DYMOA (91.8%) (minimum accélération longitudinale, maximum vitesse de rotation…).

De plus, il est possible d'étudier les performances de nos modèles pour séparer les événements des incidents en s'intéressant à la matrice de confusion (Figure 6).

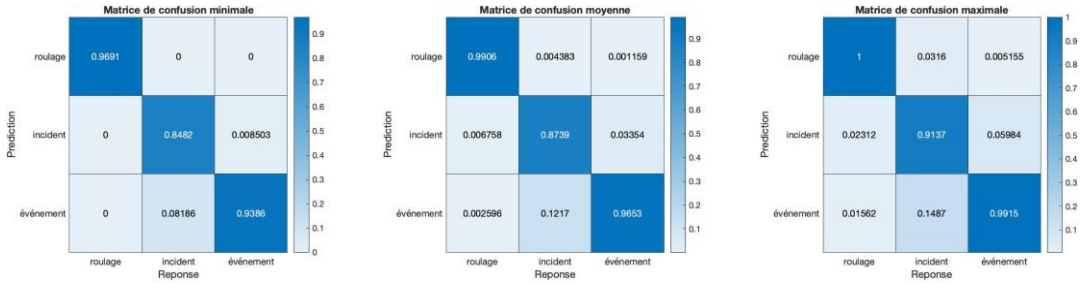

Figure 6 : Matrices de confusion

On peut voir ainsi que les observations de roulage sont très souvent bien classées, à 99% en moyenne. Les événements et les incidents sont en moyenne bien classés à respectivement 96% et 87% et en moyenne, 3% des incidents sont classés en événements.

La méthode développée dans l'étude basée sur les algorithmes d'apprentissage supervisé fournit des performances en accord avec les techniques de dépassement de seuils utilisées dans la campagne d'expérimentation de DYMOA. Pour continuer l'étude de la comparaison de méthode il a été décidé d'entraîner un arbre de décision avec les mêmes variables que celui utilisé dans le projet DYMOA. Sur la figure 7, on projette les 4 variables utilisées pour le classement originel des labels.

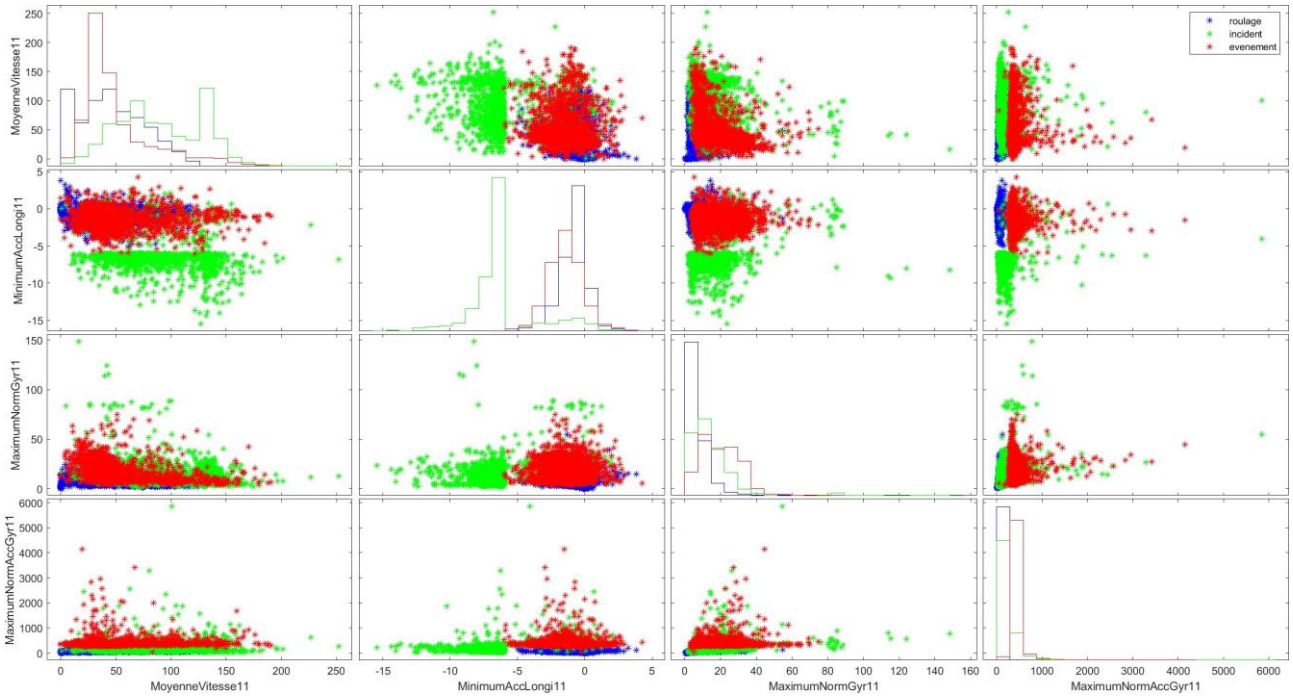

Figure 7 : Projections des observations dans un espace à 4 dimensions utilisé dans la méthode de dépassement des seuils du projet DYMOA

En figure 8 est présenté un arbre de décision entraîné avec ces mêmes variables. En utilisant toutes les données, on peut observer que le seuil de la variable " minimum d'accélération longitudinale" (MinimumAccLongi) est très proche de celui qui a été choisi dans DYMOA, soit −6 m/s².

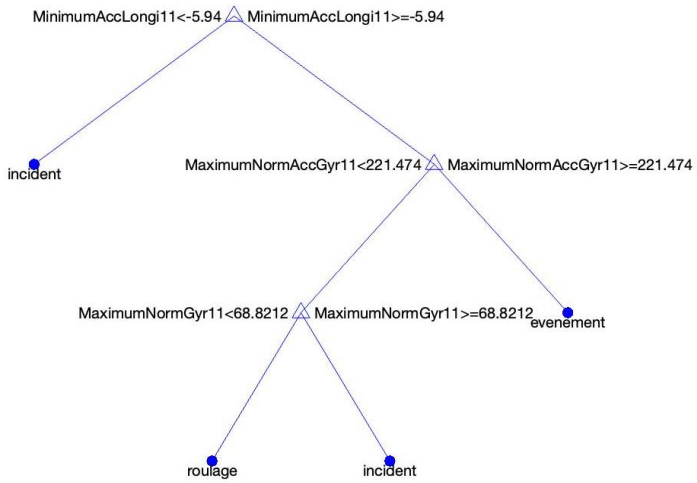

Figure 8 : Détail de l'arbre de décision entraîné avec la méthode de l'apprentissage supervisé

L'analyse des erreurs de classification notamment pour les catégories incident et évènement a fait apparaître des erreurs de labellisation. En effet les erreurs de labellisation réduiraient la

performance du modèle d'apprentissage supervisé. Par ailleurs, il n'est pas toujours évident de labelliser une situation de conduite donnée et l'erreur humaine de labellisation peut alors apparaître. En analysant les vidéos à disposition nous nous sommes aperçus que la catégorie "Incident" provenait du dépassement de seuil de 2 situations de conduite complètement différentes :

Une manœuvre d'urgence : le dépassement de seuil a lieu lorsque le conducteur effectue une manœuvre d'évitement d'une situation d'accident.

Une conduite dynamique : Le dépassement du seuil est lié à une forte sollicitation de la moto mais le conducteur ne semble pas surpris et ne fait face à aucun danger. On peut qualifier sa conduite de "sportive".

En effectuant un labélisation des données (dont notamment celle des erreurs présenter dans la matrice de confusion en figure 6) on pourrait espérer une amélioration des performances des algorithmes de machine learning. L'étude réalisée dans le Work Package 5 (WP5) se donne comme objectif d'étudier la mise à jour et l'optimisation des seuils de déclenchement lors de ces différentes situations incidents/événements.

2.4. Bilan

Cette première partie a montré que l'usage des techniques de machine learning est approprié pour la classification d'événements/incidents lors de l'usage du 2RM. La technique mise en place est relativement générique et très couramment utilisée dans le monde de l'apprentissage supervisé. Le succès de cette méthodologie dépend énormément de la qualité et du volume des données labellisées ainsi que de la capacité à évaluer des modèles appropriés aux caractéristiques du problème. Des performances de l'ordre de 94% ont pu être mesurées avec des modèles de type SVM et arbre de décision sur l'ensemble des données DYMOA. Ces modèles arrivent en très peu de temps et en utilisant peu de données à classifier les incidents et événements avec de bonnes performances. Ces performances sont comparables avec celles des techniques plus classiques utilisant la détection par dépassement de seuils. Par ailleurs une sélection des variables pertinentes a pu mettre en évidence que l'information pouvait être condensée dans une sélection de 12 variables descriptives.

L'interprétabilité de ces modèles de machine learning a permis d'identifier des variables qui caractérisent au mieux une situation d'incident ou d'événement. On retrouve notamment avec le modèle de l'arbre de décision des estimations de valeur de seuils similaires avec ceux utilisés lors du projet DYMOA et qui avaient été déterminées avec des campagnes d'essais.

Ces modèles de machine learning ont également des limites. La qualité de labellisation des observations est très importante pour les résultats finaux à la fin de la phase d'apprentissage. Ces modèles ont plus de mal à classifier la catégorie "incident" et peuvent les confondre avec la catégorie "événement". L'analyse des erreurs de classification et l'analyse des vidéos a permis de montrer que cette différence pouvait être subtile et des erreurs pouvaient exister dans la labellisation en fonction de la définition utilisée.

3. Partie 2 : Localisation des zones à risque par accumulation d'événements et incidents détectés en deux-roues motorisé

3.1. Présentation de la méthodologie d'identification d'une zone à risque par analyse statistique

L'objectif de cette partie est de pouvoir identifier des zones à risque pour l'usager du 2RM en analysant les zones géographiques d'accumulation des événement/incidents à la proximité d'un cas d'accident répertorié dans le fichier national des accidents corporels de la circulation dit "Fichier BAAC"1. La figure 9 présente une visualisation des accidents BAAC de 2RM en France sur la période de 2016 à 2018 dans les secteurs de l'étude comportant les départements : Eure, Seine-Maritime, Bouches-du-Rhône et Hérault.

On filtre les données des 2RM avec les variables ci-dessous :

- Numéro de l'accident BAAC (Num_Acc)
- Latitude et longitude de l'accident BAAC (lat et long)
- Département dans lequel a eu lieu l'accident (dep)
- Numéro de véhicule de l'accident (num_veh) car plusieurs véhicules peuvent être impliqués
- Catégorie du véhicule (catv) appartenant à la classe des 2RM :
- Cyclomoteur <50cm3
- Scooter < 50 cm3
- Motocyclette > 50 cm3 et <= 125 cm3
- Scooter > 50 cm3 et <= 125 cm3
- Motocyclette > 125 cm3
- Scooter > 125 cm3
- Catégorie d'usager du véhicule (catu) :
- Conducteur
- Passager

[1] Administré par l'Observatoire national interministériel de la sécurité routière ONISR, disponible en libre accès sur data.gouv.fr

Figure 9 : A - Visualisation des accidents 2RM répertoriés dans le fichier BAAC présents dans le secteur de l'étude; B - Zoom sur les zones de l'étude

Une fois les accidents corporels 2RM sélectionnés à partir du fichier BAAC on souhaite identifier les zones d'étude définies par la présence à proximité (distance maximale de 200 mètres) d'un ou plusieurs événement(s)/incident(s) apparu lors de l'usage d'un 2RM. On utilisera les événements/incidents mesurés lors de l'expérimentation du projet DYMOA (qui ont fait l'objet de l'étude dans la partie 1) et les zones à risques identifiés par Liberty Rider via agrégation d'événements. La figure 10 présente les différentes étapes permettant l'identification d'une zone d'accumulation commune.

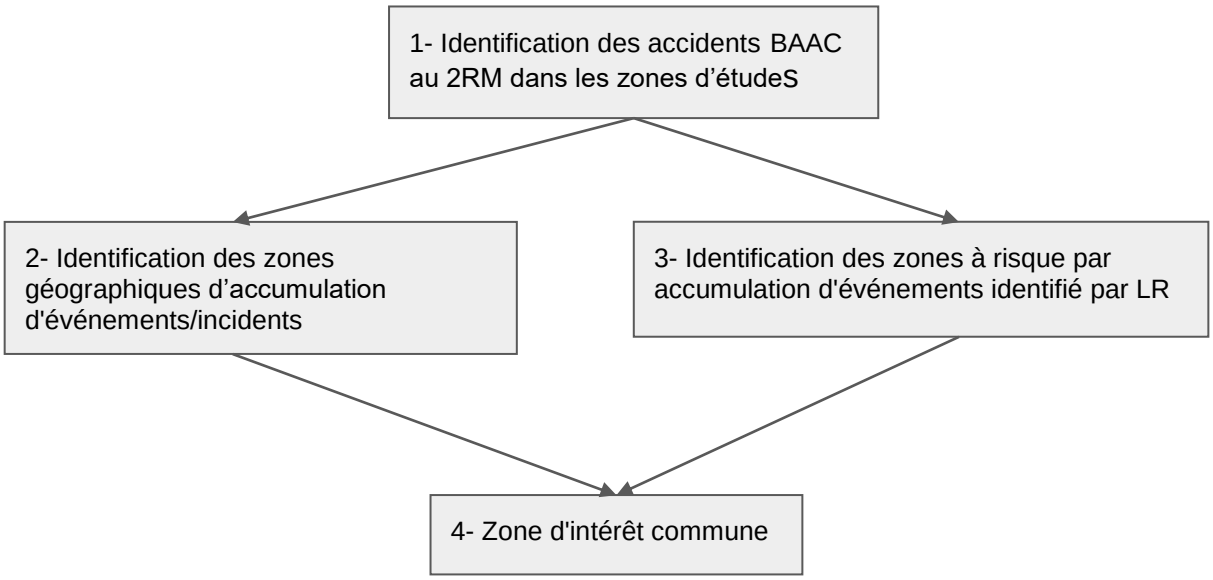

Figure 10 : Présentation des étapes d'agrégation et d'identification des zones d'accumulation

La sélection de la zone d'étude se fait en agrégeant les données d'incidents/événements dans une région circulaire d'un diamètre D, centrée sur un cas d'accident issu du fichier BAAC.

3.2. Résultats

Le tableau 3 présente le nombre de zones d'accumulation communes en fonction de la distance de la zone. On notera qu'au-dessus de 100 mètres il est difficilement envisageable de considérer qu'il s'agisse de la même zone d'infrastructure routière. Pour la suite de l'étude, on se limitera à une distance de 100m. En dessous d'une distance de 50 m le volume des zones détectées chute brutalement et représente moins de 2% du volume total des zones d'accumulation communes identifiées.

Tableau 3 : Variation du nombre de zone d'accumulation trouvée en fonction de la distance

Distance maximale	Nombre de zones communes identifiées	Pourcentage cumulé
10 m	16	0.4%
50 m	31	0.8%
100 m	1800	50%
200 m	3500	100%

La figure 11 représente les zones communes avec une présence d'accumulation des incidents et événements dans un rayon maximal de 100 m. On notera que l'on retrouve en priorité ces zones dans les grandes villes des secteurs d'étude. Cela était prévisible à cause de la forte fréquentation de ses infrastructures routières par rapport à des endroits plus reculés et donc moins fréquentés. Le biais de sélection des zones d'accumulation est donc à prendre en considération lors de l'usage de cette méthode. Il sera nécessaire de normaliser les zones d'accumulation par la fréquentation et l'usage lors ce que l'on voudra tirer des conclusions sur le caractère à risque de la zone étudiée ou lors de l'établissement du lien de causalité entre incidents et accidents.

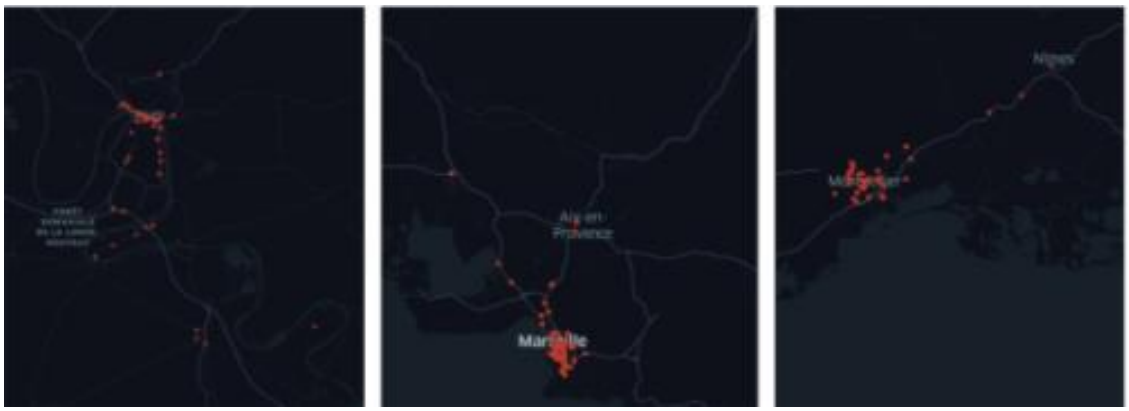

Figure 11 : Identification des zones d'étude communes dans les secteur d'étude

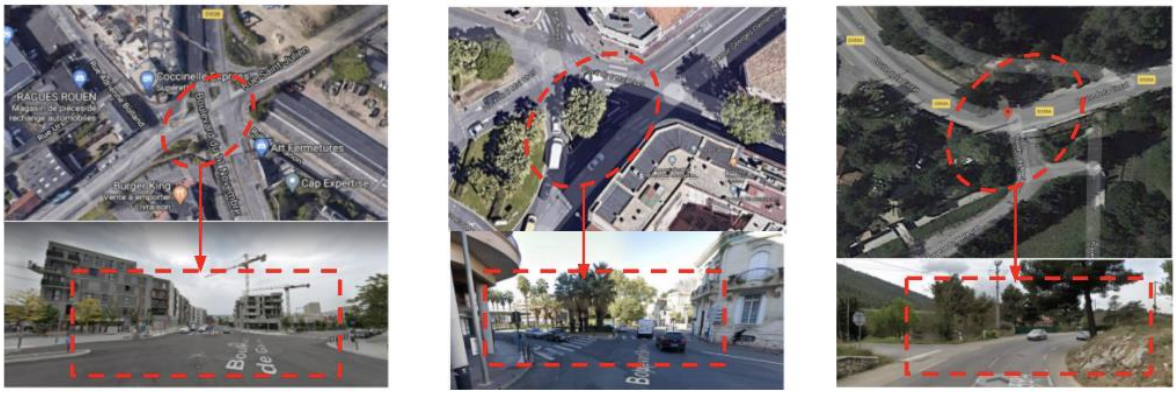

Figure 12 : Vue détaillée (Google Street Map) de 3 zones d'étude communes identifiées

La zone d'étude illustrée par la figure 12-A est une intersection en "X" (4 voies de circulation). D'après les informations sur l'accident BAAC, l'accident a eu lieu en plein jour et en agglomération. Une motocyclette et un véhicule léger sont entrés en collision par le côté sur une chaussée. Des événements/incidents ont pu être mesurés à proximité de cette zone. D'après les images fournies par Google Street Map, la qualité de la chaussée a l'air a priori tout à fait convenable. On peut émettre l'hypothèse que ces incidents et événements sont dus à l'interaction entre l'usager de 2RM et les autres usagers lors d'une circulation de trafic dense.

La zone d'étude illustrée par la figure 12-B est une intersection complexe à plusieurs voies de circulation. D'après les informations sur l'accident BAAC, l'accident a impliqué une motocyclette et un véhicule léger. Ils sont entrés en collision par le côté sur une chaussée. Des événements ont pu être mesurés à proximité de cette zone. Malheureusement les données à disposition ne permettent pas d'établir des liens de causalité entre incidents et accidents. Par ailleurs, la présence d'intersection et d'une voie à sens unique pourrait être à l'origine d'interactions complexes entre les usagers.

La zone d'étude illustrée par la figure 12-C est une entrée de virage. D'après les informations sur l'accident BAAC, l'accident a impliqué un véhicule 2RM lors d'un dépassement de véhicule. Des événements et incidents ont pu être mesurés à proximité de cette zone. Lors de l'entrée dans un virage il n'est pas rare de constater un freinage de l'usager de 2RM afin d'adapter son allure. On peut noter que la visibilité de l'entrée du virage est réduite à cause d'un accotement avec la présence de roche. Il est très difficile de comprendre avec les données à disposition le lien entre l'accident et la zone d'accumulation d'incidents.

3.3. Bilan

Cette partie montre que la méthode utilisée permet d'identifier des zones potentiellement à risque en considérant l'accumulation d'incidents et d'événements à proximité d'endroits où il y a eu un accident corporel identifié dans les fichiers BAAC. La technique d'identification de zone d'accumulation semble efficace et la distance a un effet direct sur le nombre de cas identifiés dans cette zone cible. Il semble judicieux ne pas dépasser une distance de 100 m, voire même 50 mètres en agglomération afin de ne pas comptabiliser des événements qui ne se sont pas déroulés sur la même infrastructure.

L'analyse de ces zones n'a pas permis d'établir de liens évidents entre les incidents et accidents. Plusieurs interprétations peuvent être faites concernant les limites de cette étude :

- Présence d'un biais de sélection dépendant du volume de passages de la flotte instrumentée sur le réseau routier
- Nécessité de normaliser par rapport à la fréquentation afin de pouvoir conclure sur le caractère zone à risque de la zone d'étude

- Manque d’information de contexte rendant le travail d’identification de lien de causalité difficile
- Nécessite un volume d’usage de 2RM important afin de pouvoir avoir des réponses statistiquement valides

Malgré les limitations évoquées ci-dessus, cette méthode peut être utilisée afin de mettre en lumière des zones d'intérêt qui pourront être analysées plus en profondeur avec l’expertise humaine. En effet l’identification de zone à risque et l’explication de liens de causalité entre incidents et accidents nécessite une démarche spécifique détaillée dans le Work Package 4 (WP4) s'appuyant impérativement sur informations contextuelles.

4. Conclusions

La première étude présentée dans ce rapport a montré que la méthode d'apprentissage supervisé (machine learning) est tout à fait adaptée à la classification et l'identification de situations à risque pour l'usager de 2RM. La classification des incidents/événements basées sur l'expérimentation du projet DYMOA a rendu possible l'entraînement et l'évaluation de plusieurs modèles de type machine learning atteignant des performances allant jusqu'à 94% de précision. Une attention particulière a été portée sur le choix et la sélection des types de modèles ainsi que des variables utilisées afin d'obtenir les meilleures performances au vu du volume et du type d'observations mises à disposition. Par ailleurs, l'étude a permis de montrer que l'interprétation des modèles de classification permet de retrouver de grandes similitudes avec la méthode de détection par seuil dynamique développée précédemment dans le projet DYMOA. Toutefois, des limitations ont pu apparaître au niveau de l'optimisation des performances. En effet, nous avons pu constater qu'il n'était pas toujours facile de labelliser et caractériser une situation de forte sollicitation en conduite sportive et une situation générée dans une situation de "presque accident" demandant ainsi à l'usager de 2RM de réaliser une manœuvre d'urgence. Les performances que l'on peut espérer par ces techniques d'apprentissage supervisées en dépendent grandement.

La deuxième étude présentée a permis l'identification de zones d'accumulation d'événements/d'incidents lors de l'usage du 2RM à partir du fichier BAAC, des incidents DYMOA et des zones identifiées par Liberty Rider. Ces zones d'accumulation correspondent potentiellement à des zones à risque pour le 2RM. Des limitations sont apparues quant à l'étude des liens de causalité qui pourraient expliquer le caractère à risque de la zone. Chaque situation complexe nécessite le regard d'un expert humain et une étude détaillée de la scène afin de dresser le bilan qui caractérise ou pas cette zone en termes de risque. Des éléments comme la mise à disposition de la vidéo embarquée sont d'une très grande utilité pour la compréhension de la situation.

Afin de conclure sur cette étude, nous pouvons retenir que les techniques de machine learning sont des outils prometteurs pour analyser des situations complexes d'interaction de l'usager 2RM. Elles présentent de nombreux avantages pour mettre en lumière des zones à risque qui pourront être analysées plus en profondeur avec l'expertise humaine afin d'établir un lien potentiel de causalité entre incidents et accidents.

5. Perspectives

Une poursuite logique de cette étude serait d'optimiser la performance de la méthode d'apprentissage appliquée à la classification de situations de "presque accident", situations dans lesquelles le motard fait une manœuvre d'urgence pour éviter un potentiel accident. Pour réaliser ce travail, il faudrait reprendre l'ensemble de la labellisation en appliquant avec rigueur la définition de situation de "presque accident" et ensuite ré-entraîner le modèle. Une autre suite possible serait d'utiliser les méthodes de l'apprentissage supervisées afin de classifier la cause et le type de manœuvre effectuée dans une situation d'événement / incident comme le freinage fort, le freinage d'urgence et l'évitement. Cela permettrait d'avoir une information supplémentaire aidant à la contextualisation de la situation.

Une utilisation des méthodes de deep learning pour la compréhension du contexte serait intéressante également dans de nombreuses autres études et analyses sur l'interaction de l'usager du 2RM avec l'infrastructure routière et les autres usagers. Pour cela, les données vidéo enregistrées dans le projet DYMOA pourraient être exploitées. Des techniques d'apprentissage supervisé de type Deep Learning développées au cours des dernières années ont démontré leur pertinence pour la compréhension et la segmentation d'une image. Lors de la détection d'un évènement/incident, l'analyse de l'image de la scène par ces méthodes permettrait de récupérer des informations de contexte tel que le nombre et type d'usager(s) autour, la densité du trafic, le type de la route et la présence d'un potentiel obstacle ou d'un défaut de la route. Enfin l'analyse du flux vidéo tout au long de l'usage du 2RM pourrait également être considéré comme un signal d'entrée, et corrélé aux autres signaux enregistrés des EDR et offrirait de nouvelles pistes d'améliorations des performances pour la classification de situations d'incidents et évènements. Il faut cependant vérifier la faisabilité d'une telle étude qui dépend de la qualité des vidéos, peut-être insuffisante dans le projet DYMOA, puisque cette qualité a été volontairement abaissée pour des raisons essentiellement juridiques.

Les perspectives pour l'identification de zones à risque pour l'usager de 2RM sont nombreuses. Afin de valider la performance et l'intérêt de la méthode, il serait intéressant d'identifier des zones d'intérêt en amont de l'étude afin de vérifier et analyser l'usage du 2RM dans cette zone d'étude. D'un point de vue statistique, il sera nécessaire d'avoir des renseignements sur la densité du trafic afin de pouvoir normaliser les résultats et ainsi pouvoir identifier des zones potentiellement à risque pour l'usager du 2RM avant de les analyser en détail. L'aspect temps réel pourrait également être intéressant pour acquérir la connaissance de l'évolution et de la rétention d'une zone à risque au cours du temps. L'effet de la saisonnalité et les variations des conditions météorologiques pourraient également être pris en compte dans ce type d'analyse. De plus, des techniques avancées d'apprentissage supervisées en classification d'image pourraient être utilisées afin d'analyser l'infrastructure du réseau routier par images satellites. En effet, avec l'augmentation de la précision des images satellites et la fréquence des prises de vue, il serait intéressant de savoir quel type de donnée on peut en extraire afin d'identifier des zones à risque (présence d'un obstacle sur la route, dégradation de la chaussée, etc). Par ailleurs, le type d'étude présentée dans ce rapport pourrait être généralisé à d'autres types de mobilité comme la voiture et les mobilités douces du type vélo et trottinette. Cela permettrait l'étude et la compréhension des similitudes et différences entre type de véhicule sur l'identification d'une zone à risque.

Enfin des applications industrielles pourraient voir le jour avec notamment la possibilité de fournir aux collectivités des outils permettant de cartographier ces zones à risque en temps réel. Les acteurs industriels pourraient également concevoir des systèmes d'alerte et de prévention des zones à risque pour l'usager de 2RM.

6. Références

[1] https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins

[2] https://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support

[3] https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision