ENSAE – Spécialisation Actuariat

Mémoire présenté devant l'ENSAE

pour l'obtention du diplôme d'Actuaire ENSAE

et l'admission à l'Institut des Actuaires

le 2 Novembre 2015

Par : Lison GRAPPIN

Titre: Space-time modelling of roads inherent risk using telematics data

Confidentialité : □ NON ▣ OUI (Durée : □ 1 an ▣ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présent du jury de l'Institut des Actuaires :

Signature :

Entreprise :

*Nom :*AXA Global P&C

Signature :

Directeur de mémoire en entreprise :

Nom : Philéas CONDEMINE

Signature :

Membres présents du jury de l'ENSAE :

Xavier MILHAUD

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise :

Secrétariat :

Bibliothèque :

Signature du candidat :

ENSAE ParisTech, 3 Avenue Pierre Larousse, 92245 Malakoff cedex

Résumé

Les objets connectés sont en train de progressivement révolutionner l'assurance non-vie. Les dispositifs de géolocalisation satellite et de "tracking" d'activité inclus dansce type d'objets permettent en effet maintenant aux assureurs d'acquérir une compréhension spatio-temporelle du risque porté par leurs assurés. Les données télématiques, issues de projets pilotes, sont pour le moment surtout utilisées pour créer des indicateurs de comportement de conduite susceptibles d'avoir un impact sur les risques d'accident qui débouchent sur des offres d'assurance automobile dites "pay-as-you-drive" et "pay-how-you-drive". Or, certains chercheurs ont démontré qu'au-delà du comportement du conducteur, les risques d'accident étaient aussi fortement liés aux circonstances spatio-temporelles de conduite (météo, intensité du trafic routier, pente de la route, revêtement, etc.). Certaines périodes temporelles et certains lieux seraient intrinsèquement plus risqués pour les automobilistes que d'autres.

Ce mémoire a donc pour but de montrer que les conditions spatio-temporelles de conduite doivent être mieux prises en compte par les assureurs pour, d'une part améliorer leurs analyses de risque automobile et leur gestion de sinistres et d'autre part, participer au renforcement de la sécurité routière. L'objectif final de cette étude est de créer un score de risque matérialisant le risque inhérent à chaque route en fonction de ses caractéristiques spatio-temporelles à partir des données télématiques.

Pour réaliser cette étude, nous avons utilisé les données de fréquence de sinistres et les données télématiques récoltées par l'entité Suisse d'AXA, AXA-Winterthur, qui a lancé un pilote d'offre télématique à destination des particuliers en 2013. L'étude a été réalisée en trois phases complémentaires. Tout d'abord, nous avons montré que la fréquence journalière de sinistres automobiles pouvait être en très grande partie expliquée par la météo et la temporalité. En particulier, la proportion d'accidents par conducteur augmente de 27% les jours de neige, alors qu'elle diminue de 25% les dimanches. Un important travail de valorisation statistique des données télématiques a ensuite été effectué. Celles-ci ont été traitées et enrichies pour créer des indicateurs spatio-temporels de conduite pertinents. La procédure utilisée ainsi que les problèmes pratiques auxquels un actuaire pourrait être confronté lors de l'analyse de ce type de données sont exposés en détail dans ce mémoire. Enfin, nous avons prouvé grâce à ces données que d'une route à l'autre, le risque pris par le conducteur pouvait être intrinsèquement très différent selon l'environnement géographique de la route ou encore le trafic enregistré sur celle-ci.

Les résultats de telles analyses pourraient à l'avenir venir alimenter une application mobile d'assistance à la conduite indiquant au conducteur le risque qu'il prend *ad hoc* en choisissant tel ou tel itinéraire, et en proposant éventuellement un itinéraire moins risqué. Ces applications pourraient permettre aux assureurs de devenir des acteurs à part entière des politiques d'amélioration de la sécurité routière.

Mots-clés : accidents de voiture, fréquence de sinistres, objets connectés, données télématiques, pay-as-you-drive, pay-how-you-drive, assurance "usage-based", données externes, big data, geocoding, routing, API, analyse spatio-temporelle, Gradient Boosting Machines

Abstract

Connected devices are currently deeply changing property and casualty insurance. Tracking devices indeed enable insurers to better qualify customers risk from a space-time point of view. Telematics data gathered from pilot projects are for the moment mostly used to derive behavioural features which may affect car crash risk. They are then used to feed "usage-based" motor insurance offers such as "pay-as-you-drive" and "pay-how-you-drive" products. They nevertheless still remain under-exploited with regard to space-time driving circumstances. However, some transportation researchers proved car crash risk was not only caused by drivers risky behaviour, but also by space-time driving circumstances such as the weather, the lighting, the traffic flow, the slope of the road, etc. According to these studies, specific time slots and locations would be inherently more risky than other ones.

The overall goal of this thesis is to prove that space-time circumstances must be taken into account by insurers to better monitor car crash risk and thus participate to road safety enhancement. The purpose is to finally create a scoring model that would assign to each road a measure quantifying its inherent risk thanks to telematics data.

To carry out this study, we had access to claims count and telematics data collected by AXA-Winterthur, the Swiss AXA entity, which launched a telematics pilot project in 2013 designed for individuals. Analyses were performed in three stages. First, we proved that daily claims count can be largely explained by weather and temporal features. Notably, claims count per driver increases by 27% on snowy days, while it decreases by 25% on Sundays. Then, a specific procedure was designed to deal with telematics data and enrich them to make them statistically valuable. The data processing techniques used are detailed in the report in order to highlight some of the troubles an actuary may face while industrializing telematics motor risk analyses. Finally, telematics data were exploited to show that car crash risk may be inherently very different from one road to another according to its geographical specificities or the traffic flow intensity recorded on it.

The outcomes of such risk assessments would for instance feed a driving assistance application displaying warning messages for each road of a route, and proposing alternative less risky roads to drivers, thus making insurers active stakeholders in road safety enhancement.

Key words : car crashes, claims frequency, connected devices, telematics, pay-as-you-drive, pay-how-you-drive, usage-based insurance, external data, big data, geocoding, routing, API, space-time analysis, Gradient Boosting Machines

Note de synthèse

L'objectif de ce mémoire est de montrer que les conditions spatio-temporelles de conduite doivent être mieux prises en compte dans les analyses de risque d'accident de voiture. En effet, les études actuelles, tant en assurance que dans le cadre des actions de prévention pour améliorer la sécurité routière, se concentrent surtout sur les caractéristiques des conducteurs à risques (caractéristiques individuelles et comportments dangereux au volant). Cependant, de l'expérience de n'importe quel conducteur, il existe des endroits ou des moments qui s'avèrent intrinsèquement plus risqués que d'autres, quelque soit son comportement au volant. On peut par exemple penser à une route de montagne sinueuse et étroite, à une route verglacée ou encore à une période de conduite sous la neige. Notre but est donc ici de statistiquement mettre en évidence l'apport des facteurs spatio-temporels aux analyses de risque automobile.

Contexte de l'étude

Pour mener à bien ce type d'analyses statistiques, il est nécessaire d'utiliser des données datées et précisément localisées géographiquement. Or, les objets connectés, qui sont en train de progressivement révolutionner les techniques actuarielles classiques d'assurance non-vie, répondent à ce besoin. En effet, les dispositifs de géolocalisation satellite et de "tracking" d'activité contenus dans ces objets permettent maintenant aux assureurs de récolter en temps-réel des données géolocalisées spécifiques à chaque assuré, et de retracer ainsi l'usage qu'il fait de ses biens assurés. En assurance automobile, ces objets connectés récoltent des données dites "télématiques".

Pour réaliser cette étude, nous avons ainsi utilisé les données télématiques récoltées par l'entité Suisse d'AXA, AXA-Winterthur, qui a lancé un pilote d'offre télématique à destination des particuliers en 2013, ainsi que ses données de sinistres. Etudier des données suisses est d'autant plus intéressant du point de vue de notre étude que la Suisse est un pays à la topographie complexe (des régions de haute montagne cotoient des vallées très urbanisées) et soumis à de fortes pressions en terme de trafic routier.

L'étude a été réalisée en deux temps. Nous avons tout d'abord analysé la fréquence journalière de sinistres d'un point de vue purement temporel en intégrant des données externes de météo et de trafic routier. Nous avons ensuite étudié cherché à caractériser le risque intrinsèque des routes à l'aide de leurs caractéristiques géographiques principalement. Nous avons pour cela créé un modèle de "scoring" des routes qui attribue à chacune une mesure quantifiant son risque intrinsèque à partir des données télématiques.

La fréquence journalière de sinistres automobiles peut être expliquée en grande partie par des déterminants météorologiques et temporels

Pour modéliser la fréquence journalière de sinistres en collision, nous l'avons tout d'abord étudiée comme une série temporelle afin d'en tirer des indicateurs de saisonnalité. Nous l'avons ensuite modélisée par un modèle de comptage intégrant à la fois les indicateurs de saisonalité créés précédemment et les circonstances météorologiques et de trafic enregistrées chaque jour. Nous avons ainsi pu dégager quelques résultats intéressants :

- Le trafic routier doit être interprété comme une variable d'exposition : si le nombre de véhicules présents sur une route augmente, le nombre d'accidents de voiture et de collisions potentiels augmente mécaniquement.
- La proportion d'accidents de voiture par conducteur diminue de 25% les dimanches et de près de 19% les jours de vacances. En revanche, elle augmente de plus de 27% les jours de neige.

Néanmoins, ces effets sont susceptibles de changer de manière importante d'une région à l'autre, d'une ville à l'autre, voire d'une route à l'autre. Ces analyses doivent donc être complétées par des études prenant en compte des paramètres géographiques qui puissent être localisées au niveau de chaque route.

Comment donner de la valeur statistique à des données télématiques ?

Les données télématiques sont très précieuses pour réaliser ce type d'étude car elles rassemblent à la fois des données spatiales (positions GPS) et temporelles (heure et date d'enregistrement de la position). Cependant, les offres d'assurance télématiques étant toujours en phase de test, la méthode de collecte de ces données n'est pas encore optimale. Il n'est pour l'instant pas possible d'intégrer directement ce type de données dans des modèles statistiques. Elles doivent être nettoyées, complétées et enrichies par des données externes pour pouvoir vraiment apporter une valeur ajoutée dans des études de risque.

Il n'existe pour l'instant pas encore de méthode de traitement des données télématiques en assurance qui soit complètement établie. C'est pourquoi nous avons développé, pour mener à bien nos analyses, notre propre méthode d'enrichissement des données télématiques. Les principales étapes de cette procédure sont les suivantes :

- Les données doivent d'abord être complétées grâce à un algorithme de *routing* qui permet de retrouver, à partir d'une suite de points, l'itinéraire emprunté par l'utilisateur et le relier à une suite de routes existantes dans le réseau routier étudié. L'ampleur du trafic routier passant spécifiquement sur chaque route est alors estimée.
- Les données de sinistre doivent être "reverse géocodées", c'est-à-dire que le nom de la route sur laquelle chaque sinistre a été enregistré est déduit de sa position GPS.
- Ces deux bases de données sont ensuite fusionnées afin de lier, pour chaque route, sa mesure d'exposition (le trafic routier) à sa mesure de risque (le nombre de sinistres localisés sur cette route).
- Enfin, cette base de données fusionnée est enrichie grâce à des données externes provenant de réseaux "open-data" répertoriant des données de météo, des données topographiques ainsi que des données caractérisant l'environnement immédiat de la route.

Les actuaires doivent considérer cette phase de traitement et d'enrichissement des données télématiques comme la phase la plus exigeante et la plus chronophage dans la réalisation d'études de risques intégrant ce type de données. Nous avons cependant bon espoir que celle-ci pourra s'accélérer dans les prochaines années grâce au développement des processus de traitement des "big data".

Quantification du risque d'accident intrinsèque des routes

Nous avons choisi de matéraliser le risque intrinsèque associé à une route par le nombre d'accidents corporels qui ont été enregistrés sur celle-ci par les autorités suisses. C'est ainsi la prédiction de ce comptage fournie par notre modèle statistique, dérivé de techniques de machine learning, qui constitue le score de risque associé à la route.

Tout d'abord, pour prouver que les indicateurs spatio-temporels créé lors de l'enrichissement des données télématiques sont pertinents pour prédire le risque intrinsèque des routes, nous avons procédé graduellement. Nous avons ajouté étape par étape les variables topographiques, météorologiques et enfin environnementales en tant que variables explicatives dans notre modèle, et étudié la variation en termes de pouvoir prédictif du modèle induite par ces ajouts. Nous avons ainsi pu prouver que les variables environnementales permettent d'améliorer bien plus sensiblement le pouvoir prédictif du modèle que les variables météorologiques.

Ce modèle de "scoring" nous a ainsi permis de montrer que le risque inhérent aux routes varie d'une route à l'autre selon ses caractéristiques et son environnement géographique immédiat :

- Plus le trafic routier enregistré sur la route est important, plus le score de risque est élevé. Par exemple, pour les routes comptant plus de 400 passages de voitures dans notre base de données, doubler le trafic revient à augmenter le nombre d'accidents corporels de 150%.
- Plus une route est longue, plus son score de risque intrinsèque est élevé, car l'automobiliste est exposé plus longtemps à un risque de collision.
- Les routes situées dans des zones de haute montagne, c'est-à-dire au-dessus de 3500 mètres, comptent en moyenne deux fois plus d'accidents corporels que les autres routes.
- Les routes principales (majoritairement les autoroutes), enregistrent en moyenne deux fois plus d'accidents corporels que les routes tertiaires reliant des localités de petite et moyenne taille.

Applications possibles et futurs développements

Les études spatio-temporelles réalisées dans le cadre de ce mémoire, et particulièrement celles se basant sur les données télématiques, en sont encore à l'état d'ébauche. Les techniques utilisées mériteraient d'être utilisées sur des données collectées sur des périodes plus longues et enrichies de manière encore plus précise. Mais il ne fait nul doute que les offres télématiques seront bientôt proposées à une plus grande échelle, ce qui permettra une amélioration des processus de collecte et de traitement de ces données.

A l'avenir, les résultats de telles analyses pourraient venir alimenter une application mobile d'assistance à la conduite indiquant au conducteur le risque qu'il prend *ad hoc* en choisissant tel ou tel itinéraire, et en proposant éventuellement un itinéraire moins risqué. Ces études spatio-temporelles pourraient aussi être utilisées par les régulateur pour repérer des "points noirs" routiers associés à des scores de risques intrinsèques très élevés. Ceci leur permettrait ainsi mieux allouer leurs budgets dédiés à la maintenance des routes

Ces applications pourraient permettre aux assureurs de devenir des acteurs à part entière des politiques d'amélioration de la sécurité routière.

Executive summary

The overall goal of this thesis is to show that space-time driving circumstances should be better taken into account in car crash risk analyses. Indeed, car crash risk studies carried out by insurers and road safety enhancement organisations focus on drivers individual characteristics and risky driving behaviours. Nevertheless, any driver will say that he knows some time slots and places that are more risky to drive than others : a mountainous winding road, a frozen road or driving by snowy weather. Our purpose is thus to statistically prove that space-time features are relevant to model car crash risk.

Context

In order to achieve these analyses, properly dated and geo-located data are necessary. Hopefully, connected devices, that are currently greatly changing property and casualty landscape, meet this need. Indeed, tracking devices now enable insurers to real-time collect individual geo-located data and thus derive how the goods they insure are used by the customer. In motor insurance, such devices are called "telematics".

To carry out this study, we used claims count and telematics data collected by AXA-Winterthur, the Swiss AXA entity, which launched a telematics pilot project in 2013 designed for individuals. Study Swiss data was even more adapted to our purpose as Switzerland has a harsh topography (mountainous regions are neighbouring heavily urbanized valleys) and has to cope with heavy traffic flow.

This space-time car crash risk study was performed in two stages. First, daily claims count was studied from a pure temporal point of view, and was modelled using external weather and traffic data. Second, we tried to qualify roads inherent risk using mainly geographical data. We ended up creating a scoring model that assigns to each road a measure quantifying its inherent risk thanks to telematics data.

Daily claims count can be explained to a large extent by weather and time features

Daily collision claims count was first examined as a time series, to derive seasonal features. We then modelled it using a count data model that included the seasonal features derived from the time series analysis, weather and traffic data as covariates. It enabled us to retrieve some relevant time and weather features to model and predict the daily temporal evolution of motor collision claims count :

- Traffic flow intensity has to be taken into account as an exposure variable : the more vehicles are travelling on the roads, the automatically greater is the number of collision claims recorded.
- The proportion of daily collision claims per driver decreases by 25% on Sundays, and by nearly 19% on public holidays, but it raises by more that 27% on snowy days.

Nevertheless, these effects may greatly change from one region, locality or even road to another, according to their geographical specificities. Thus, claims count analyses must not only be carried out at a country level, but also at the local level, and especially at the road-level.

How to make telematics data statistically valuable ?

Telematics data seem optimal to carry out such space-time car crash risk analyses, as they gather both space and time features while real-time recording GPS trip positions. Nevertheless, their collection process is still in the early stages, and the data recorded cannot be directly included in statistical models as things stand. They must be properly processed and enriched using external data to actually bring value to statistical analyses.

For the moment, there is no reliable method available to process such data. We thus had to design our own data processing method in order to achieve our study. It goes through the following steps :

- Trips data has to be completed using a routing algorithm that derives the existing roads of the network on which drivers travelled, and retrieve the traffic that went through them.
- Claims data has to be reverse geocoded, that is located on an existing road of the network from a GPS location. A claims count can thus be attached to each road.
- Trips and claims data has to be merged in order to link, for each road, its exposure parameter (traffic count) with its risk measure (claims count).
- External open data (weather, topography and environment data) were finally used to associate to each road its specific surroundings and topographic characteristics, as well as weather features.

Actuaries must consider this data processing stage as one of the most challenging and time-consuming phase when integrating telematics data in their motor risk analyses. We however hope that we will be able to make these processes quicker in the coming years, thanks to big data processing techniques development.

Roads inherent risk quantification

We chose to measure road inherent risk by the number of bodily injury crashes reported on it by Swiss police force. Thus, the risk "score" we sought to create corresponds to the this bodily injury crashes count predicted by our machine-learning based statistical model.

First, in order to prove that space-time features are relevant to model car crash risk on a road level, we gradually included topographic, weather and surroundings features as covariates in our statistical models. At each step, we computed a prediction quality measure and finally showed that it improved at each step. It enabled us to demonstrate that surroundings features greater improve model prediction power than weather ones.

This risk scoring model enabled us to establish that roads inherent risk may vary greatly from one road to another according to its own geographical characteristics :

- The greater traffic flow is recorded on the road, the more numerous bodily injury crashes it counts. For instance, when the telematics traffic count exceeds roughly 400 cars, doubling the traffic count on the road increases the number of bodily injury crashes located on it by nearly 150%.
- When looking at roads of more than 1 kilometre length, the longer they are, the more numerous bodily injury claims count are recorded on them.
- Roads located in high mountainous localities (beyond 3500 meters high) count twice as many bodily injury crashes as the other roads.
- Primary roads, that is mostly highways, record twice as many crashes as tertiary roads, that is roads linking medium-sized towns, and smaller types of roads.

Potential further applications and improvements

The analyses performed on telematics data in this report are still at the draft stage. Indeed, their collection process and external enrichment still need to be improved. It requires nevertheless more accurate databases and claims telematics data recorded on longer periods, but we have no doubt that it will become possible in the coming years thanks to the development of such products on a broader level.

In the future, risk assessments using telematics data may be beneficial to drivers. They would indeed for instance feed a driving assistance application displaying warning messages for each road of a route, and proposing alternative less risky roads to drivers. Furthermore, such analyses may also be beneficial to regulators : by pointing out roads that are inherently very risky, insurers can help governments to better allocate their road maintenance budgets and give priority to so-called "black-spots" locations.

These applications will certainly promote insurers as active stakeholders in road safety enhancement.

Acknowledgements

I would like to express my sincere gratitude to :

- Philéas CONDEMINE, my supervisor, for his great patience, motivation, enthusiasm and strong support throughout my work. I would like to thank him gratefully for his useful suggestions concerning statistical and actuarial matters and for his helpful hints in critical steps of my work.

- Guillaume GORGE, Head of AXA Global P&C Technical Excellence, who provided me an opportunity to join his team as an intern. I am very grateful to him for his guidance and advice.

- Christoph BUSER and Huan LIU, R& D actuaries at AXA-Winterthur in Zürich, for their patience and constant support throughout my internship. I would like to sincerely thank them for providing the telematics data I worked on and for the useful insights they gave me while analysing my results.

- Stéphane SOULIER, Senior Retail Data Scientist at AXA Global P&C, who provided insight and coding expertise that considerably assisted and accelerated my research work.

- Eliane VO from AXA Group Solutions, and Ahmed TOUATI from Data Innovation Lab for their great help while enriching telematics data.

- Godefroy BEAUVALLET, Head of AXA Ecosystem team, for his challenging suggestions concerning the future of telematics, and his knowledge about connected devices and API.

- Pierre RINDER, Selim RABOUDI and Siyuan DAI, my fellow interns, for the stimulating discussions, and for all the fun we have had in the last six months.

- AXA Global P& C IT team and Data Innovation Lab employees, who gave me access to research facilities and helped me deal with big data matters.

- All the members of the Data Science and Retail teams, for their warm welcoming.

- Xavier MILHAUD, my academic director and former head of the Actuarial Science program at ENSAE, for his reviews and advice.

Table of contents

Introduction

Road injuries caused circa 1.3 million deaths in the world in 2012, ranking them ninth among the main causes of deaths reported on that year, despite "being predictable and largely preventable", according to the World Health Organisation. Enhancing road safety must thus be considered by regulators and governments as a major public health issue. Most of European countries thus created synergies between road users associations, governmental agencies and police force to deal with this issue. Their actions generally take the form of road safety prevention campaigns warning about *risky driving behaviour* (under alcohol or drugs influence, fast driving, etc.), laws banning such driving attitudes and police enforcements. Despite these measures road fatalities records remain high, which potentially means that these prevention measures are not efficient enough, or do not target all the possible car crash risk factors.

Actually, some transportation researchers argue that, beyond behavioural causes, car crash risk is partly explained by space-time circumstances, such as weather, traffic flow intensity or even road geometry. According to these studies, specific time slots and locations (roads or crossings for instance) would be inherently more risky than other ones, whatever the driver behaviour.

Nevertheless, to carry out such risk studies at a large scale, it is crucial to have access to dated and geolocated driving and crash reports data. This is where the insurers can bring a clear added to value to road safety enhancement. Indeed, they have been collecting motor claims data for years and are currently beginning to collect telematics data, which enable them to acquire an accurate space-time comprehension of car crash risk.

The overall goal of this thesis is thus to prove, thanks to claims frequency and telematics data, that space-time circumstances are relevant to analyse car crash risk, and that they must be better taken into account by actuaries when performing their risk studies. The final purpose of this study is to create a scoring model that will associate to each road a measure quantifying the inherent risk bore by it.

To carry out this study, we had access to claims count and telematics data collected by AXA-Winterthur, the Swiss AXA entity, which launched a telematics pilot project in 2013. After having detailed why telematics data are relevant to study car crash risk, we will introduce the main characteristics of car crash risk in Switzerland. The actuarial risk analysis of car crash risk was then performed in three phases. First, a temporal analysis was performed on collision daily claims count, aiming at quantifying the impact of harsh weather and traffic conditions on daily claims count records. Second, a specific procedure was designed to make telematics data valuable for statistical studies. These data were processed and enriched to create relevant space-time features. The data processing techniques are detailed in the report, in order to highlight some of the troubles an actuary may face while industrializing telematics motor risk analyses. Third, telematics data were exploited to create a roads inherent risk scoring model.

PART 1

Monitor car crash risks using telematics data

According to the European Commission, circa 30 000 people died on the roads of the European Union in 2011 [1], which amounts to the population of a middle town. Moreover, for each road fatality, around 4 permanently disabling injuries, 8 severe injuries and 50 minor injuries are reported. As road safety is a major social issue, the European Union targeted in its latest strategic plan to lessen road fatalities count by at least 10 000 by 2020. This figure dropped to 26 000 in 2013 thanks partly to this plan, but it remains a high record that needs to decline.

Road prevention is the major channel used by governments to tackle this issue, but it is not sufficient. Private stakeholders should also be involved in reducing the risks, and among them especially insurers who have a broad overview and experience of motor vehicle risks. Motor vehicle insurers have indeed been collecting individual data and motor claims data for a long time mostly for pricing purposes. But this data is also widely used to better understand the risks undertaken by the company and monitor the loss ratio. Through these actuarial and statistical analyses, insurers derive in fact useful insights about car crash risk factors. While a decade ago such analyses were most of the time carried out at a yearly aggregate level, increasing data storage capacities and machine learning development boosted insurance real-time data collection through connected devices, such as motor "telematics" insurance products. Analyses can now be performed on an individual level with spatio-temporal and real-time dimensions that can make car crash risks studies far more accurate.

In this chapter, we will first define the "telematics" offer launched almost recently by several insurers, then look at well-known car crash risk factors pointed out by transportation researchers. Finally, we will explain why telematics data are relevant in this context.

1.1 What are "telematics" ?

"Telematics" is a generic word referring to connected devices for vehicles that register real-time data, especially GPS coordinates, technical measures such as braking, wheels rotation, etc. The telematics device that registers data usually takes the form of a smart-phone application or a sort of black box plugged directly into the car. It is now mostly related to motor vehicle insurance through the "pay-how-you-drive" products that use "telematics" technology to track customers' usage of vehicle.

1.1.1 Usage-Based insurance business model

According to the insurance credibility theory, pure premium must be derived from an additive mixture of collective risk experience and individual risk carried by the customer himself. When the customer underwrites its insurance contract, the insurer does not have any piece of information about its current and future behaviour. He can only deduce its potential behaviour toward risk according to objective characteristics and collective experience he has from previous customers who share the same characteristics. Concerning motor vehicle insurance, the underwriter will base its pricing on individual characteristics such as age, gender (outside Europe), profession, etc. [1] and on vehicle characteristics such as its type, its size, its construction date, etc. Then, the insurer will gain experience with this specific customer as soon as it registers claims, and he will modulate the customer's pure premium to better match his behaviour and assess his own risk. The insurer will gradually partition its portfolio using the experience he gains from the customer.

[1] In Switzerland, gender and nationality features are still allowed to be used in insurance pricing models.

But this pricing method is limited as it does not take the actual behaviour of the customer into account. Indeed, we derive this behaviour only from the claims the customer reports, but we do not know actually if he drives everyday or only twice a month, if he drives aggressively or safely, if he drives at night on mountain roads or in the traffic jam, etc. **"Usage-Based Insurance"** (UBI) is a specific type of insurance product that adapts pure premium to the actual behaviour of the customer, using mostly real-time data [25]. There are two major types of UBI insurance products for motor vehicles :

- **"Pay-As-You-Drive" (PAYD)** or "Mile-based" insurance : pure premium calculation is based on the number of kilometres driven. The principle is that a customer who drives less is less exposed to traffic and car crash risk and thus less risky from the insurer point of view. His former pure premium (which means "traditionally" calculated) will thus decrease in order to adapt to his actual driving usage.
- **"Pay-How-You-Drive" (PHYD)** insurance : pure premium calculation is based on the *way* of driving of the customer. More precisely, the telematics device plugged in the car real-time records all technical parameters related to each trip, such as GPS coordinates or acceleration and braking parameters. Thus the insurer knows precisely where, when and how the customer drives and derives from it information about its way of driving (aggressive, safe, slow, etc.). Pure premium is thus individualized to match precisely the behaviour of the customer, and may be discounted by as much as 30% if the customer's way of driving is considered as "safe" [7].

Actually, both the insurer and the insured benefit from Usage-Based Insurance products, and especially telematics devices. Indeed, the insurer gets a lot of accurate information about his customer and can thus better assess his risk and avoid anti-selection bias [2], while the insured will get a discount on his pure premium if he drives safely.

Moreover, insurers generally grant an additional discount to customers who simply agree to plug a telematics device in their car. We can wonder why they grant such an additional discount even before collecting any data about their customer : it has in fact to do with social psychology. According to Jeremy Bentham social theories [6], just realizing that you are watched is sufficient to behave safely and according to the rules. Bentham applied this principle while theorizing his *Panopticon* jail scheme, in which a unique watchman is needed whatever the number of prisoners there are in. The prison building is designed as a circle whose center is a watching tower, with the unique watchman in it. Prisoners are in individual cells around it and thus know they could be watched at any time by both the watchman and the other prisoners. Bentham explained that, as soon as the prisoner knows he is being watched, he acts with respect for the rules. Another example of this phenomenon could be found in behavioural economics experiences [5]: an honesty box was settled near the coffee machine in a university, with the rule that while pouring oneself a coffee, students had to give some coins, but no amount was clearly indicated. It was shown that on days when a photo of eyes was pinned above the honesty box, students tend to pay nearly three times as much for their drinks than on other days. This backs up the fact that telematics drivers would drive generally more safely than usual because they are being watched, and even more because they are financially incited to do it (the pure premium discount). Telematics devices thus participate to road safety improvement, which is one of its worth advantages. From the insurer point of view, it has also the advantage to reduce *a priori* claims frequency.

1.1.2 A brief history of telematics insurance

The American insurance firm *Progressive* was the first to launch a pay-how-you-drive telematics pilot program in 1998, but stopped it in 2001 because of too high information technology costs, and a too tiny data storage capacity. Moreover, synergies between car manufacturer, telecoms companies and insurance companies had to be strengthened to improve the project. After several years of research and development, *Progressive* finally launched the actual first insurance telematics product called "Snapshot" in 2011. *Progressive* was a forerunner and thus has been a leader in this market for a long time. Its competitors, including AXA, launched similar offers in the following years and developed it worldwide. This specific insurance market is growing exponentially, driven partly by the connected devices and "big data" revolution. AXA

[2] "Anti-selection" phenomenon appears when an insurer has a not enough segmented pricing strategy : "good" risk carriers tend to pay too much compared to their actual risk and will turn to the competition, while "bad" risk carriers tend to pay less than their actual risk. This leads to attract "bad" risk carriers, instead of "good" risk carriers, which is contradictory with the insurer's goal.

initiated its telematics service offerings first in Italy and Switzerland as pilot entities in 2008, and Ireland in 2013. The program was then extended to several entities across Europe, especially Switzerland with the "Drive Recorder" on which this study is based, in United Kingdom with the "Drivology" system, in Spain, etc. and represents now circa 200 000 insurance policies across Europe.

Nevertheless the telematics market is still in the early stages and some issues often raised by the media need to be tackled to enable larger expansion. Among them, privacy protection is the most critical and discussed one, as in other connected devices markets such as activity tracking watches. Indeed, telematics devices register all trips driven by the customer which may be seen as highly intrusive, especially in Europe or France where privacy protection is strictly supervised by authorities. Privacy and data protection matters are still a major brake for growth. Customers unions also fear that telematics data could be used against their interest as it could lead to a premium raise if the driver behaviour is flagged as hazardous by the insurer algorithm. This hypothesis can in fact be challenged, as drivers who don't behave safely, knowing they will be watched, will generally not subscribe to such an offer. Telematics insurance offers are initially meant to reward *safe* behaviours. Insurers, data storage firms, regulators and customers are still negotiating those matters in order to enable this promising market to grow under control.

Figure 1.1: *Progressive* "Snapshot" device

1.2 Overview of car crash risk factors

According to the World Health Organization (WHO) [24], about 1.24 million people die each year worldwide due to road traffic crashes, especially young people aged between 15 and 29 years. Moreover about 20 to 50 million people suffer non-fatality injuries often leading to disability, involving huge economic losses and costs. The WHO points out that this issue has been mostly "neglected from the global health agenda for many years, despite being predictable and largely preventable". As road safety has become a critical public health matter worldwide , numerous specific research studies about car crash risk factors have been published.

Phase		**Risk factors**		
Phase	Action	Human	Vehicles & equipment	Environment
Pre-crash	Crash prevention	- Information - Attitudes - Impairment - Police enforcement	- *Road-worthiness* - Lightning - Braking - Handling - Speed management	- *Road-design* and road layout - *Speed limits* - Pedestrian facilities
Crash	Injury prevention during the crash	- Use of restraints - Impairment	- Occupant restraints roadside objects - Crash protective design of the car	- Crash-protective - Other safety devices
Post-crash	Life sustaining	- First-aid skill - Access to medics	- Ease of access - Fire risk	- Rescue facilities - Traffic congestion

Table 1.1: Haddon's matrix

One of the first researcher to have studied specifically road traffic injuries in a public policy framework is William Haddon Jr. in 1980 [15]. As the leader of the National Highway Traffic Safety Administration and of the Insurance Institute for Highway Safety in the USA, he made numerous contributions to the injury control field thanks to his research activities. His most well-known conceptual work, now called the "Haddon's matrix" (see table 1.1), aims at understanding how road traffic injuries occur and thus identifying key risk factors before developing public policy strategies to reduce those risks. Based on the work of J.E.

Gordon and J.J. Gibson [14], he divides time in three phases : the "prevention period", corresponding to the pre-crash stage ; the "interaction period", corresponding to the crash stage, and the "salvage period", corresponding to the post-crash stage. These stages define the rows of the matrix, which columns depict three major risks factors : human, vehicle and equipment, and environment (physical, socio-economical and legal). This matrix shows an overview of risks factor that need to be taken into account while studying car crash data. In our study, we will focus on non-human pre-crash risk factors (Vehicle & Equipment and Environment columns), as our goal is to assess physical road safety characteristics [3].

1.2.1 Human risk factors : driver's characteristics and behaviour

What the WHO calls "predictable and largely preventable" car crash causes are mostly related to the driver's characteristics and behaviour. Road users are well aware of them and road safety prevention campaigns generally focus on them. Though our study aims at focusing only on non-human risk factors, we depict here briefly the most important human risk factors enlightened by the WHO [24].

Young adults and especially males are far more involved in car crashes than other population categories. Indeed, young adults (from 15 to 44 years old) account for circa 60% of global road traffic fatalities, and among them three-quarters are specifically due to male drivers. A well-known statistics state that young male under 25 are three times as likely to be killed in a car crash as young female. Young drivers lack of experience and more hazardous behaviour are at stake here. But these figures need to be linked with specific behaviours known as key car crash risk factors.
First, high speed is increasing the likelihood of a crash occurring, and furthermore of severe crash occurring, typically when exceeding maximum speed limits. Second, drink-driving is known for sharply increasing car crash likelihood above a blood alcohol concentration of 0.4 gram per blood litre. WHO states that "enforcing sobriety checkpoints and random breath testing can lead to reductions in alcohol-related crashes of about 20%" [24]. Medicinal and recreational drugs that impair driving performance are also often cited, even if the direct relationship between dose level of drugs and increasing car crash likelihood is a complex matter as they are often mixed use alcohol drinking. Third, motor-cycle helmets, sea-belt and child restraints can reduce the risk of death or severe injury in a car crash by almost 50 % according to the WHO. Finally, distracted driving due to the use of mobile phones is a growing cause of car crash, leading to longer reaction times and shorter following distance.

1.2.2 Driving environment risk factors

Research studies focusing on driving environment risk factors can be classified into two categories : "aggregate" studies that try to explain and model claims frequency, and "disaggregate" analyses that consider specifically pre-crash circumstances and that use real-time data as far as possible. As these works handle with space-time circumstances of car crashes, they are most of the time restrained to tiny areas such as a particular highway portion and carried out on a short time period. Nevertheless, they are worth reading since they provide useful insights on potential risk factors involved in road traffic crashes.

Traffic flow as an exposure parameter
Road traffic crashes generally fall into one of the three common categories depicted below :

- Lane-change related crash : it occurs when the driver tries to change lane and collide with another vehicle or roadside object.
- Junction related crash : it occurs at crossroads and lead to rear-end collision and angle or side impacts
- Crashes involving other road users or objects : it involves pedestrians, cyclists or animals

From this crashes classification, we can derive that crashes occurrence and types are heavily linked to traffic flow and local environment. Indeed, a lane-change related crash is more likely to happen on a highway with heavy traffic flow, whereas crashes involving other road users or objects are more likely to occur in urban (when involving pedestrians) or countryside areas (when involving animals) with low traffic flow. It obviously shows that traffic flow should be considered as a critical *exposure* parameter (see chapter 3).

[3] Italicized words emphasize the risk factors we will spotlight through our analyses.

For instance, Abdel-Aty and Radwan are exploring ways to model crashes frequency using traffic flow in their article entitled "Modelling traffic accident occurrence and involvment" [3]. Their study focuses on a highway in Florida (USA) and aims at linking crashes frequency with traffic flow while controlling driver characteristics. In order to do it, they use two datasets arising from Florida Departments of Highway Safety and Transportation that contains crashes reports on one side and traffic flow and road geometry description on the other side. They compare then two widely used count data models, a Poisson regression and a Negative Binomial regression. They finally chose the Negative Binomial one as their data is highly over-dispersed [4]. They draw specific attention on the use of "exposure variables" : they include for example the logarithm of AADT (Annual Average Daily Traffic Data) per lane of the highway as an explanatory variable and show that it was heavily significant to explain crashes count. Multiplying ADDT per lane by 2.7 raises the likelihood of accidents by 86%.

Road design

Abdel-Aty and Radwan [3] were not the first researchers to show that AADT has a strong impact on the likelihood of road traffic crashes, but their work was pioneering in the inclusion of road geometry components in crashes frequency modelling. Using the phrase "road geometry variables", they mean road length, road surface, degree of horizontal curvature, shoulder median width, lane width over number of lanes ratio and a dummy indicating if the road is located in a urban or countryside area. They use a sort of clustered approach to fit their model : they divide the highway they are focusing on in segments of homogeneous horizontal curvature and width. Thus, for each segment, they count the number of crashes that occurred on it, which corresponds to the response variable of their model, and use the segment length as an additional exposure variable. Undeniably, the longer is the length of the road segment, the more numerous crashes are likely to occur on this segment. The results show how strong is the influence of road geometry on crash count :

- An increase in the degree of horizontal curvature of the road section leads to an raise of 13% in the the number of crashes.
- An increase in the shoulder width has a negative impact on the number of crashes (lessening of 12%), as it hinders roadside object hitting.
- An increase in the lane width over number of lanes ratio (which amounts to the width per lane) has a strong negative impact on the crashes count, as it reduces the crash count by circa 44%.
- If the road section is located in a urban area, it increases crash count recorded by 35%.

Nevertheless, Abdel-Aty and Radwan draw attention on the fact that these features and their impact need to be revised for each driver category. The accident database they use contains the driver's individual characteristics involved in each crash. They fit therefore separate Negative Binomial models for each category of drivers (male/female drivers, young/middle/old drivers), and study the changes in coefficients and significance involved compared to the general model. They demonstrate that, according to the specific population at stake, the most significant variables are not the same and the effects of each variable on the crash count is different. For example, an increase in shoulder width has a greater impact on crash count when looking at old drivers, and an increase in width per lane has a greater impact on crash count when looking at female drivers. This article points out that, when analysing road geometry risk factors, and even all non-human risk factors, it is critical to control model fitting by individual characteristics when available in order to better understand the phenomena. Though, one of the limits of this model is that it focuses only on crashes count and does not take real-time data into account, overseeing specific car crash circumstances. Furthermore, it focuses only on a unique 227 km length major highway in Florida, which is not enough to draw generalities. We will thus work on a larger scale of data, focusing on a lot of roads and cluster them according to their environment specificities.

Another article by Ahmed and Abdel-Aty [4] explores the road geometry risk factor in a mountainous context, namely an interstate highway segment in the American state of Colorado with a downward slope of more than 7%. The starting assumption of their work is that mountainous roads are generally known to be more hazardous than on other roads because of harsh geometric features such as corners, sharp downward slope, etc. Moreover, they acknowledge that weather conditions in a mountainous environment can change greatly within a day and may be dramatically impacted by high elevation.

[4] We will be confronted with the same problem while choosing the most appropriate claims temporal model (see chapter 3).

Figure 1.2: Alpine winding mountainous road

Their goal is to assess interactions between road geometry, climate and crashes frequency, and to spot specific hazardous road segments. They work on homogeneous road sections in the same way as the previous article. Using successively a hierarchical Bayesian Poisson model, a random effect model and a spatial model, they prove that, if located in a mountainous environment, curvature and median width have a mere effect on crash frequency. When comparing these results to the findings of Abdel-Aty and Radwan we derive that the same road geometry features can have very different impacts on crash frequency according to the surroundings of the road we look at. Finally, their geographic approach (working with road sections) enables risk ranking. Indeed, in the last part of their article, they rank each road section location according to the hazard it carries in terms of crash likelihood using a Bayesian score. This geographical scoring technique may be useful for our purpose.

Weather and lightning conditions

A cornerstone article was published in 2003 by Golob and Recker [13] : it is one of the first article that explores the crossed influences of traffic flow, weather and lightning conditions on car crashes liability. They qualify the types of traffic accidents occurring on urban freeways in South California according to the traffic flow and the weather and lightning conditions that prevailed half an hour before the car crash reported in their database. The originality of their approach lies in their attempt to work with near real-time data. In order to get the actual circumstances that prevailed half an hour before the car crash, they use Induction Loop Detectors (IDL). An IDL is a sensor embedded in the road that records each car passage on the road and the time associated. This system is widely used on highways to automatically record daily traffic flow. Golob and Recker create their data features by matching these IDL records with timely weather and lighting data.

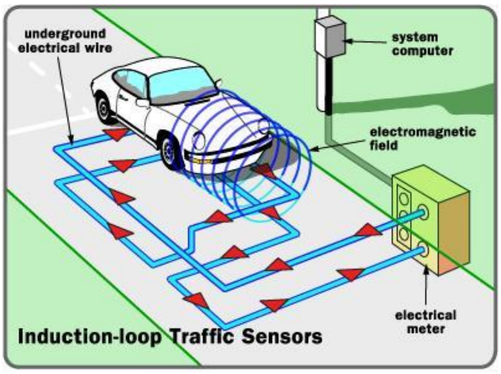


Figure 1.3: Induction Loop Detector system

In order to get more interpretable results, they reduce the number of modalities for each type of variables : weather data is summed up to dry or wet ; lightning to dawn, daylight, dark lighted or dark unlighted ; car crash severity to injury or property damage only ; accident depiction to off-road, rear-end or object hitting. Golob and Recker apply then a non-parametric canonical correlation analysis on these variables in pursuance of extracting meaningful correlations. Their results are summarized figure 1.4 : severe accidents occur mostly at night by wet weather, while rear-end collisions occur mostly in traffic jams by day.

Nevertheless, this study was carried out on only 2000 car crashes in a very specific area, Southern California, which topography is relatively flat. Furthermore, one of the most important limit of this study is that it does not include actual real-time data : IDL systems do not allow individual car identification in the traffic flow count records. We thus do not know for instance what was the actual speed of the damaged car, on which lane it was situated, etc.

Coming back to Ahmed's and Abdel-Aty's [4] study on mountainous environment, it is also innovative in the sense that it includes climate features in the models fitted through a "season" parameter. They establish that crash risk during the snow season is circa 82% higher than crash risk in the dry season, all other variables remaining equal. This increased risk within the snow season may be due to the combined effect of snowy, icy and slippery pavement conditions that are made worse by the abrupt slopes of mountainous roads. It will thus be necessary to pay attention to potential interactions between weather conditions and

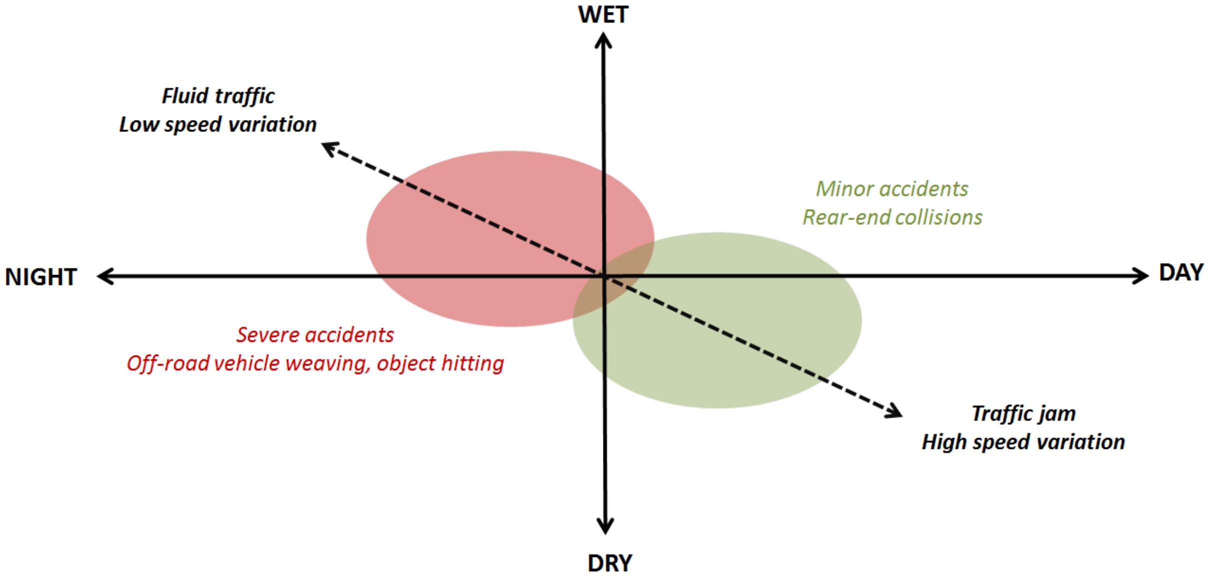


Figure 1.4: Synthesis of car crash risk factors derived from GOLOB and RECKER

road design in our further analyses. Switzerland is indeed a mountainous country (see figure 2.1).

This brief review of scientific research on car crash risk factors provided us an idea of how intricate the interactions between human, environment and equipment pre-crash risk factors (referring to Haddon's classification) can be. Besides, it emphasizes the fact that few research studies had been performed using actual real-time data, as it is far more difficult to collect or to have access to. Furthermore, all these articles focus only on restricted areas, especially major roads because data is also far more difficult to collect on secondary roads.

1.3 Why are telematics data relevant to better monitor car crash risks ?

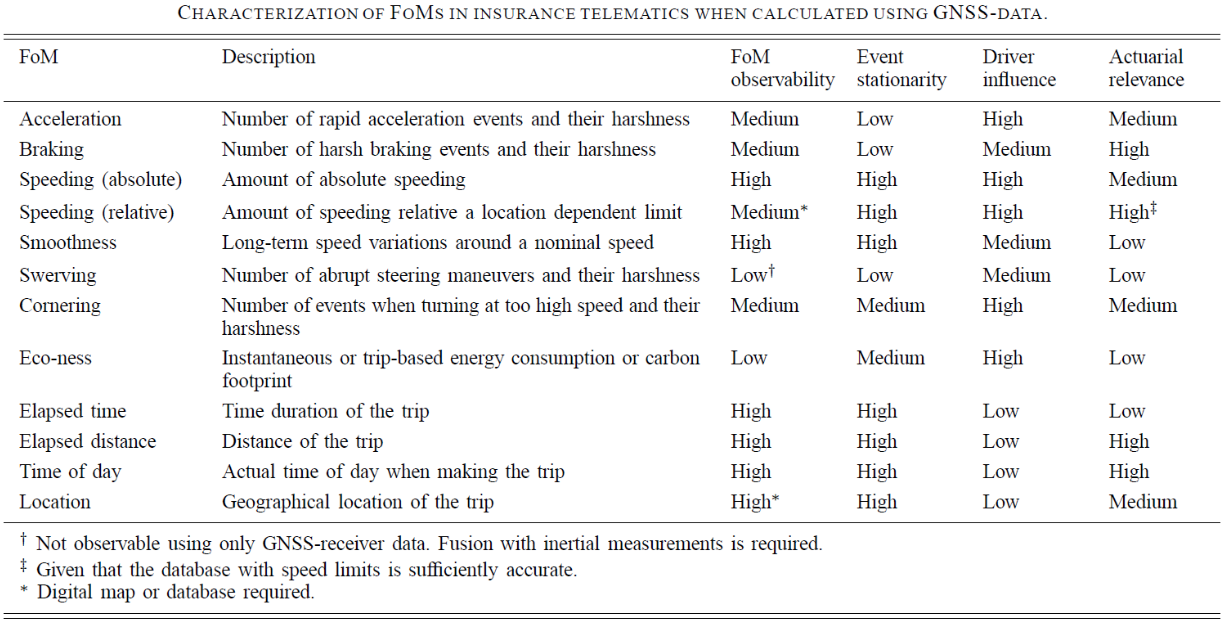
Telematics data have a competitive advantage compared to the databases studied in the research articles presented above : it is real-time collected and is directly related to a specific individual car. Moreover, as it is collected by insurers, we can link it to the driver's characteristics such as his age, health condition, place of living, etc. Indeed, telematics gather all data relevant to explore each type of pre-crash risk factors enlightened by Haddon in a unique dataset on an individual scale :

- Human risk factors : The driver's behaviour and habits can be derived, such as his usual travels, his aggressiveness while driving, his past claims ...
- Vehicle and equipment risk factors : Telematics device records all vehicle parameters such as acceleration, braking, ...
- Environment : GPS coordinates enable precise geo-location that guarantees a high coverage of the road network, and when enriched with external data, enables surroundings depiction (urban or countryside area, land-use, type of road, maximum speed limit, etc.). Moreover date and time data collection allows for matching weather and lighting conditions, traffic flow counts.

From a risk prevention point of view, telematics data are thus really valuable. Few research studies were carried out using these databases for the moment because telematics service offerings are relatively new and the data collected is owned directly by insurers and stays highly confidential. But this scientific field is growing quiet rapidly, boosted by machine learning techniques development.

Furthermore, telematics data have a strong direct interest in actuarial risk studies. Indeed, as Händel et alii stress it [16], each feature collected by the telematics device plugged in the car may be used to better price motor insurance contracts. Table 1.5 [5] enlightens the actuarial relevance of each driving feature collected by telematics devices. We note that even technical parameters such as cornering, speeding or smoothness can be employed in an actuarial analysis according to Händel. Actually, these parameters may be included in a machine learning model deriving the "way of driving" of the customer, thus inferring the risk he carries, and better classify him as "hazardous" or "safe" for example. This prevents both anti-selection and bad pricing. Insurers are currently starting to take advantage of these telematics data for their pricing models and techniques. For instance, AXA launched a *Kaggle* [6] challenge in 2014 which goal was to detect occasional driver's trips in a wide range of telematics data. The models submitted will certainly be exploited in the future to better acknowledge the actual usage of a familial car within a year for example, and thus better fit the pure premium that has to be paid for complete risk coverage. However, this contest focused only on behavioural factors (drivers' habits), but as we saw in the previous research articles, driving environment should not be overlooked. That is why we conducted our study specifically on driving circumstances, in order to prove their relevance in actuarial car crash risk studies.

Telematics data remain nevertheless almost under-exploited, both in actuarial studies and risk prevention research articles. Because of their generally massive volume and their specificities (geographical data especially), telematics data management and treatment still needs to be improved and industrialized to be plainly usable by insurers. Our study can also be seen as an attempt to rationalize telematics data treatment in order to make them statistically valuable (see chapter 5).



CHARACTERIZATION OF FOMS IN INSURANCE TELEMATICS WHEN CALCULATED USING GNSS-DATA.

FoM	Description	FoM observability	Event stationarity	Driver influence	Actuarial relevance
Acceleration	Number of rapid acceleration events and their harshness	Medium	Low	High	Medium
Braking	Number of harsh braking events and their harshness	Medium	Low	Medium	High
Speeding (absolute)	Amount of absolute speeding	High	High	High	Medium
Speeding (relative)	Amount of speeding relative a location dependent limit	Medium*	High	High	High‡
Smoothness	Long-term speed variations around a nominal speed	High	High	Medium	Low
Swerving	Number of abrupt steering maneuvers and their harshness	Low†	Low	Medium	Low
Cornering	Number of events when turning at too high speed and their harshness	Medium	Medium	High	Medium
Eco-ness	Instantaneous or trip-based energy consumption or carbon footprint	Low	Medium	High	Low
Elapsed time	Time duration of the trip	High	High	Low	Low
Elapsed distance	Distance of the trip	High	High	Low	High
Time of day	Actual time of day when making the trip	High	High	Low	High
Location	Geographical location of the trip	High*	High	Low	Medium

† Not observable using only GNSS-receiver data. Fusion with inertial measurements is required.
‡ Given that the database with speed limits is sufficiently accurate.
* Digital map or database required.

Figure 1.5: Specificities of driving features collected by telematics devices by Händel [16]

[5] "FoM" states for "figure of merit".

[6] *Kaggle* is a web platform dedicated to data science open challenges : more detailed can be found on the following website : https://www.kaggle.com/

PART 2

Car crash risk in Switzerland

As mentioned above, insurance telematics products are almost recent and have spread in Europe for only a few years. AXA Winterthur, the Swiss entity of AXA, was one of the first entities in AXA Group to launch such products : the "Crash Recorder" in 2008 and the "Drive Recorder" in 2013. The "Crash Recorder" was a pilot project that consisted in a device recording data only when crash is occurring [1]. The "Drive Recorder" is a more complex device that real-time records technical parameters for each trip driven. Our study is based on the latter recorder data coming from Swiss AXA customers who subscribed to this offer. As our goal is to use this data to better assess *environmental* car crash risk, we focus in this chapter on the car crash risk peculiarities in Switzerland.

2.1 Background information

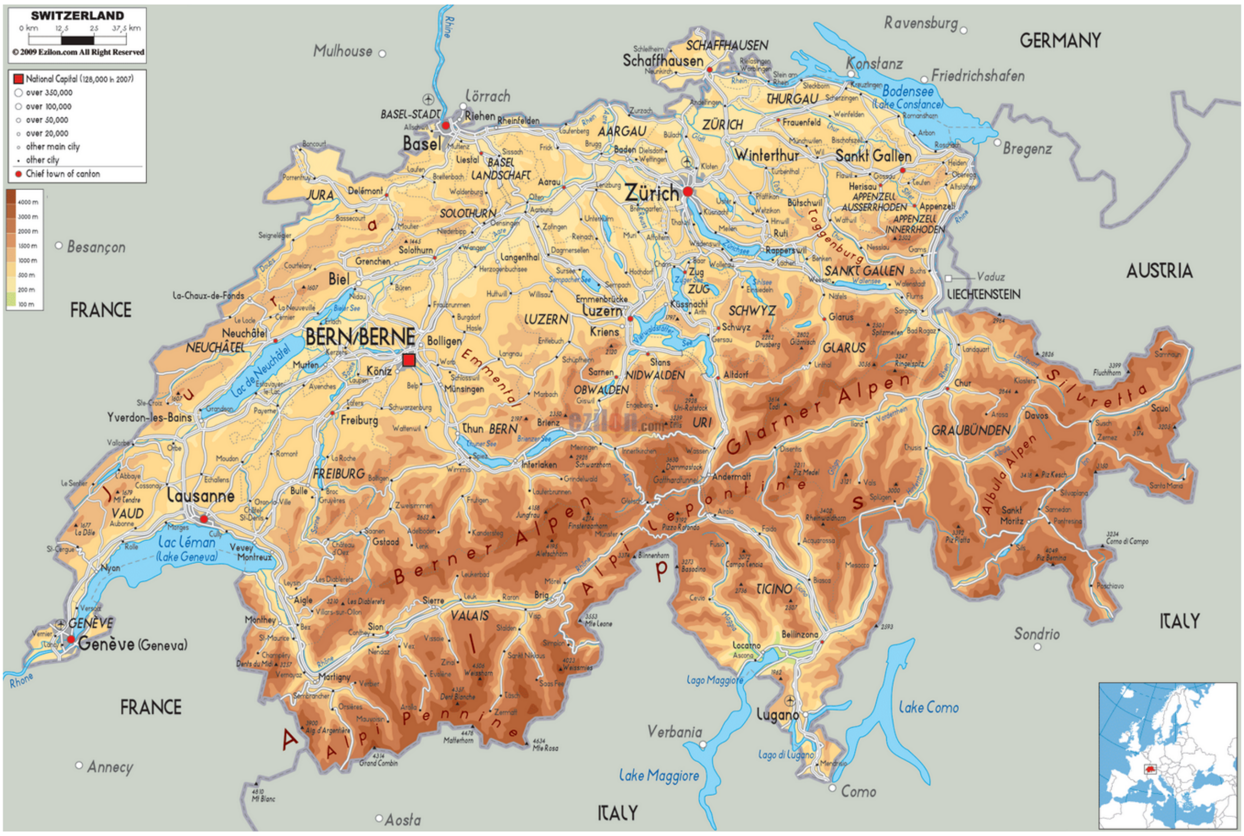


Figure 2.1: Topographic map of Switzerland

Switzerland is an 8-million inhabitants mountainous country located at the convergence of the Alps, the Jura and the Swiss Plateau, spanning an area of circa 41 000 km [2]. Joining Italy, Germany, France and Austria, it holds a central position in Europe (see Figure 2.1), and is thus confronted to a huge road traffic. According to the OECD annual report on road safety [18] and the Federal Roads Office of Switzerland (FEDRO), the road traffic in Switzerland is driven by two major tendencies : transport freight through the

[1] A "crash" recording is triggered if a set of specific parameters exceeds some technically defined thresholds : acceleration parameters for instance, pressure ...

Alps and tourism. The proportion of foreign vehicles going through Switzerland is above 70% [18] : indeed, Switzerland is well-known for its cross-border workers (circa 300 000 individuals) and is a favoured gateway for tourists to travel from the north of Europe to the south. The domestic vehicle fleet of Switzerland is moreover relatively high compared to its neighbouring countries : in 2012, there were 705 vehicules per inhabitant in Switzerland, against 650 in France and 632 in Germany. But according to the OECD [18], this does not lead in average to greater road fatalities compared to other countries : in 2012, there were 4.3 road fatalities per 100 000 inhabitants in Switzerland, against 5.8 in France and 4.4 in Germany. Furthermore, Swiss road fatalities and injuries decreased respectively by 63 % and 24% between 1990 and 2012 [18], while within the same time period, the number of domestic vehicles registered increased by nearly 50% and the number of kilometres travelled raised by 22% (see figure 2.1 [2]). This is mostly due to proactive road safety programmes initiated by the Swiss government and the Federal Roads Office during the last decades (see 2.2). Finally, the OECD estimates that road crashes costed circa 10.4 billion euros in 2009 in Switzerland, a huge amount that appeal for even more road safety prevention measures.

	1990	**2000**	**2010**
Fatalities	925	592	327
Injury crashes	23 834	23 737	19 609
Billions of kilometres travelled	49.6	55.7	62.3
Death per billion of kilometres travelled (I)	18.6	10.6	5.2
Percentage of change from 1990 in I	-	- 43%	- 72%

Table 2.1: Time evolution of Swiss road safety indicators [18]

2.2 Road safety handling in Switzerland

Many organisations are in charge of road safety at different scales (regions, local, federal) in Switzerland, due to its specific federal political scheme. Three organisations are leading the stream at the federal level and are responsible for the wider road safety improvement plans : the "Fund for road safety", the "Beratungsstelle für Unfallverhütung" (BFU) or Swiss Council for Accident Prevention and the Federal Roads Office (FEDRO). Finally, the Swiss Federal Council regulates the national road safety policy and oversees the three latter organisations.

2.2.1 National measures undertaken

Road safety has been a major political issue in Switzerland since the middle of the 2000s. At that time, the target was to decrease the number of road fatalities and severe injury crashes reported in 2000 on Swiss roads by a half by 2010. Despite tangible road safety improvements, the goal was not reached. That is why broader plans were undertaken by Swiss federal authorities to greatly impair the number of bodily injury car crashes. Among those measures, two programmes are worth mentioning : the "Strassenverkehrsgesetz" bill (SVG) and the "Via Sicura" program.

"Strassenverkehrsgesetz" bill

The SVG bill was issued in 1959 by the Swiss Federal government in order to set a global federal legal framework concerning road maintenance and road safety all across the country. Every year since, several amendments are introduced to improve road safety. Among those amendments, some are worth mentioning to better understand the broad Swiss legal framework about road safety :

- The Federal Roads Office, created in 1998, was made responsible for the security and functionality of Switzerland's motorways and main roads all across the country. Its principal missions deal with road maintenance, network watching and road safety enhancement.

[2] An *injury crash* is a car crash resulting in at least one injured or killed person.

- Several conventions between Switzerland and other European country have been passed in the 2000s to enable suing foreign drivers who violate traffic rules.
- In 2005, the maximum legal blood alcohol concentration (BAC) was decreased from 0.8 g/L to 0.5 g/L and random breath-testing was introduced. The limit for drugs has been set to zero for many years.

Via Sicura

In June 2012, the Swiss Federal Council adopted the "Via Sicura" road safety program which includes a wide range of road safety measures meant to become progressively effective till 2017. Even if no quantitative target has been set for this program, the measures undertaken are ambitious and will certainly have a strong impact on the enhancement of road safety. Here are some examples of the initiatives included in the program :

- Road safety audits are conducted on a regular time basis (see 2.4.3).
- Mandatory tests on fitness to drive after a conviction of drug or alcohol offences.
- Revocation of driver's licence for two years minimum in case of exceeding maximum speed limits and lifelong revocation in case of repeated conviction.
- Alcohol locking system [3] mandatory for excessive drunk driving offenders.
- Novice drivers as well as professional drivers are subject to a zero alcohol limit.
- Daytime running lights are mandatory for all motorised vehicles whatever the weather.
- Bicycle helmet use is mandatory on electrical bicycles above 25 km/h assistance.

2.2.2 FUSAIN : a research project dedicated to road safety enhancement

In addition to these legal measures meant to improve road safety, Switzerland supported several research programs to better understand its own car crash risk and locate specific areas where investments are necessary to enhance road safety. The "FUSAIN" program, launched between 2007 and 2009, was one of them. "FUSAIN" states for FUsion of SAfety INdicators and was co-managed by a French research institute on Transportation called the IFFSTAR and the EPFL in Lausanne, a top scientific university. The goal was to study the validity of existing collision risk indicators usually employed to assess road safety and improve them, especially by adapting their computation to be real-time done. This project was also meant to emphasize the role played by weather data on road safety : new indicators were created to better take it into account (see the "seasonality" study below [20]). Researchers carried on several case studies mainly on Swiss motorways ([9], [19], [20], [21]).

One of these case studies achieved by Pham and De Mouzon aimed at pointing out "pre-crash" circumstances on motorways [21], that may later be used to warn drivers when they encounter such a situation. Two types of crashes were scrutinized : "rear-end" crashes that often occur in high speed and heavy traffic flow, and "side-swipe" crashes that happen generally when drivers change lanes. Pham and De Mouzon spot four well-known "Safety indicators" that can be exploited to characterize a "pre-crash" situation :

- "Time To Collision" (TTC) (Hayward 1972) : time needed for two vehicles to collide if they remain on the same track and keep the same speed.
- "Potential Index for Collision with Urgent Deceleration" (Uno 2002) : index that quantifies collision likelihood in case of harsh braking performed by the leading vehicle.
- "Individual Braking Time Risk" : braking duration needed not to collide the immediate preceding vehicle in the same lane.
- "Speed Over Speed Limit" : discrepancy between actual speed and maximum speed limit.

They then computed these indicators on a specific highway junction ramp in the Vaud region, known to be risky. These indicators enabled them to score each vehicle driving on this highway section and define them as a "risky" vehicle or a "safe" one. Nevertheless computed scores are not enough robust as safety predictions change greatly from a safety indicator threshold to a slightly different one. Despite these limits, this article enlightens how harsh it is to define understandable safety thresholds while limiting the number of false warnings. The telematics database we will work on contains a variable called "event"

[3] Drivers that possess this system need to alcohol breath-test themselves each time they want to drive. If the breath-testing is positive, the car is locked and the driver cannot open it nor start it.

based on similar "safety indicators". An "event" is reported in the database each time some vehicle technical parameters exceed a certain threshold (see part 5), which means that the driver has performed a "unsafe" maneuver that could have lead to a crash (harsh braking, harsh cornering ...).

Another study focuses on meteorological data linked to traffic data [20]. Pham and Chung used Swiss Automatic Road Traffic Counts (SARTC) data (see 2.4.2) and linked it to weather data furnished by Boschung road weather stations [4]. They look at A1 and A9 motorways situated in the Vaud region. Using these datasets, they explored the validity of a "seasonal" effect on traffic : are traffic conditions dependent on the weather and the temperature in Switzerland ? They concluded that according to the weather (dry, wet or snowy), average speed of vehicles differ greatly on the motorways studied : by dry weather, speed is higher (87.9 km/h); by harsh weather (especially if it is snowy), drivers seem to be more aware of risks and drive less speedy (86.3 km/h on average). Moreover, they enlighten a substantial "seasonal" effect on traffic : the average speed during non daylight in the third quarter (autumn) is significantly higher than in the first quarter (winter) by 1.5 km/h. It thus seem that drivers adapt their driving to winter conditions because they are aware of increased road risks due to harsh weather conditions. Even if this study was carried out on a tiny portion of Swiss motorways, these insights can help us better interpret our claims frequency modelling results (see part 3) using weather data.

2.3 Characteristics of car crashes in Switzerland

Swiss federal government strongly tackles road safety matters by promoting broad action plans. But these plans cannot be effective if they are not based on a clear understanding of underlying car crash risks. Knowledge on car crash risk has been boosted in the last decades in Switzerland thanks to accurate crashes data collection and statistics reports. We introduce in this section key figures, and perform a geographical analysis of car crash risk in Switzerland.

2.3.1 Crashes data collection

The Federal Roads Office is the Swiss federal authority responsible for road infrastructure and private road transport. Since January 2011, it is in particular in charge of the collection of Swiss car crash data. A data collection platform was created to gather as much detailed statistical and geographical data as possible for each car crash reported. To be more precise, the car crashes that are fully reported in this database correspond to the *bodily injury car crashes* that involved at least one injured body, even slightly. Indeed, this database is build on a cross-checking of police reports and insurance data, and police forces intervene on a car crash scene only if an injured person has been reported [5].

The FEDRO associates to each crash a large range of features : for example, it defines several degrees of severity in bodily injuries [6], reports its GPS position, fully explains the circumstances of the accident (date, time slot, type of collision, number of vehicles/pedestrians implicated, etc.). Furthermore, since 2014, the FEDRO combines its data with hospital data to get even more details on the type of injuries caused by the car crash. It even created an internet platform presenting this dataset on a Swiss map [2]. This database is thus extremely valuable from an insurer point of view : by knowing precisely when, where and how car crashes occur in Switzerland, motor risk assessments may be far more accurate than they are today. Nevertheless, in order to protect privacy, access to this database is strictly regulated. Aggregated datasets or bare car crashes positions [7] only are available. We used both types of databases in our study over the

[4] *Boschung* is a Swiss firm that specialized in vehicle and road maintenance services and in mobile services dedicated to drivers. It especially possess a wide network of weather stations all across Switzerland to supply its driving warning platform.

[5] A mere part of the crashes reported may be linked to crashes in which nobody was injured, but it remains highly marginal. It can happen when one of the drivers involved in the crash asks for the police to come to the location, even if nobody is injured.

[6] The FEDRO defines several degrees of severity for each crash [18] : a *slight injury* corresponds to minor injuries such as superficial skin injuries ; a *serious injury* is reported when at least one person is hospitalised or not able to do its daily activity for at least 24 hours ; a *road fatality* corresponds to a death that occurs within 30 days succeeding the car crash.

[7] What we call "bar car crashes positions" corresponds to a database containing only latitude and longitude position of each car crash reported by the police between 2011 and 2013. Car crash time and date are not available, nor severity, depiction or circumstances of the crash. See section 2.4.1 for more details.

years 2011 to 2013 : aggregated data was exploited to draw a broad picture of car crashes in Switzerland in this chapter, and bare car crashes positions were used in our road risk scoring based on telematics data (see part 5).

2.3.2 Focus on road fatalities [8]

The number of road fatalities decreased sharply in Switzerland within the last 20 years, involving thus great changes in the characteristics of such road accidents [18]. First, a reduction of road fatalities has been observed in all age groups, but the decrease of the number of young people killed (aged 18 to 24) is far more sharp than in the other age groups, as it has decreased by nearly 83% between 1990 and 2010. Second, pedestrians and motorcycles are more represented in road fatalities compared to car passengers (see table 2.2), who are more protected by the vehicle itself. Nevertheless, not wearing a seatbelt increases sharply the probability to be killed as a car passenger : 56% of road fatalities between 2010 and 2012 did not wear a seatbelt. Finally, road fatalities occur more often on rural roads than on motorways or inside urban areas whatever the year (1990, 2000, or 2010) : nearly 55% of road fatalities crashes occur on rural roads.

	1990	**2000**	**2010**
Number of road fatalities	925	592	327
0-17 years	8.1%	9.1%	6.1%
18-24 years	23.1%	15.4%	11%
25-64 years	47.4%	48.1%	52%
More than 65 years	21.4%	27.2%	30.9%
Pedestrians	18%	22%	22.9%
Motorcycles/Cycles	28.3%	26.9%	32.1%
Cars	53.6%	51.2%	44.9%

Table 2.2: Percentage of road fatalities by type of road user and by age group [18]

2.3.3 What are the main causes and circumstances leading to a car crash ?

Figure 2.2: Alcohol testing on a motorway in Switzerland

One of the most publicised and well-known reason for having a car crash is the abuse of alcohol or drugs. A lot of road safety awareness campaigns focus on these risk factors, as they are more controllable than other risk factors. But actually, car crashes due to dugs or alcohol abuse represented only 10% of all injury car crashes in 2014 in Switzerland, which amounted to 1 800 crashes out of 17 800 [9]. Most of injury car crashes in Switzerland are in fact due to at least three other reasons pointed out by the OECD [18] : distraction or lack of attention especially due to the use of smart-phones while driving which caused crica 28% of injury crashes ; "inappropriate speed" which is responsible for circa 28% of road fatalities [10] ; tiredness (2% of all injury crashes according to police reporting). These figures show that human factors may play a great role in the triggering of car crashes. But at the same time, the fact that distraction represents a great part of car crashes causes advocates in a way in favour of the substantial impact of road geometry and surroundings may have on car crash risk. Indeed, on a complex-geometry road or on a heavily travelled motorway, even the mere lack of attention can lead to a crash. On the contrary, on a long and broad straight road (typically on a motorway), ability of drivers to concentrate tends to drop, leading to a raise of car crash likelihood. Road geometry may thus have an ambiguous outcome on car crash risk that we will further investigate.

[8] We point out here just orders of magnitude to have a better broad view of the car crash risk profile of Switzerland.

[9] Figures are extracted from the aggregated database delivered by the FEDRO, that concerns all bodily injury car crashes (and not only road fatalities).

[10] OECD reports that in 2010, 23% of drivers drove above the speed limits on urban roads, 31% on rural roads and 18% on motorways.

As mentioned in the previous part [13], weather and specific road environment can increase greatly car crash risks. In order to get first insights of the impact of those phenomena in Switzerland, we computed some univariate statistics on aggregated *bodily injury* car crashes data of 2013 delivered by the FEDRO [11]. Tables 2.3, 2.4 and 2.5 summarize the results. Some figures are worth mentioning :

- Bodily injury crashes seem to occur on roads with a low maximum authorized speed : more than a half of injury crashes in 2013 occurred on roads limited to 40 to 80 km/h. Crashes on this type of roads may be due to what the OECD report calls "inappropriate speed" driving, which means partly exceeding the authorized threshold.
- Bodily injury crashes seem to occur less on motorways than on secondary roads in urban and countryside areas (9.6% against 64.5% and 25.9% respectively). These figures need nevertheless to be offset by traffic exposure, phenomenon that will be carefully looked at when modelling bodily injury crashes count (see chapter 5).
- On motorways, bodily injury crashes happen more by wet weather than in average : 28.5% against 23.6% in average (all type of roads considered).
- On countryside roads, bodily injury crashes occurrence is much higher by snowy or icy weather than on average : 9.6% against 5.3% in average (all type of roads considered).
- More than 72% of bodily car crashes happened within daylight period in 2013, between 6 am and 5 pm.

Maximum speed	**10-30 km/h**	**40-50 km/h**	**60-80 km/h**	**90-120 km/h**	**Total**
Number of crashes % of total	936 5.4%	9 712 55.6%	5 538 31.7%	1 287 7.3%	17 473 100%

Table 2.3: Repartition of crashes according to the maximum authorized speed on the road in 2013

Crash location / Weather	**Dry**	**Wet**	**Snow/Ice**	*Total*
Urban area	8 290 73.6%	2 610 23.2%	362 3.2%	*11 262* *64.5%*
Countryside	3 042 67.3%	1 037 22.9%	443 9.8%	*4 522* *25.9%*
Motorway	1 088 64.4%	481 28.5%	120 7.1%	*1 689* *9.6%*
Total	*12 420* *71.1%*	*4 128* *23.6%*	*925* *5.3%*	*17 473* *100%*

Table 2.4: Car crashes repartition according to their location and the weather

Crash location / Time slot	**00-05 h**	**06-11 h**	**12-17 h**	**18-24 h**	*Total*
Urban area	666 5.9%	3410 30.3%	4 913 43.6%	2 273 20.2%	*11 262* *64.5%*
Countryside	336 7.4%	1 267 28.0%	1 944 43.0%	975 21.6%	*4 522* *25.9%*
Motorway	154 9.1%	538 31.9%	701 41.5%	296 17.5%	*1 689* *9.6%*
Total	*1 156* *6.6%*	*5 215* *29.3%*	*7 558* *43.3%*	*3 544* *20.8%*	*17 473* *100%*

Table 2.5: Car crashes repartition according to their location and the time slot within which they occurred

In the next chapter, we will model the links between daily claims frequency and some of the phenomena we pointed out here. We will also use those insights to analyse telematics data in the last part of this report.

[11] The car crashes characteristics enlightened here on the 2013 database are almost the same in the 2011 and 2012 databases.

2.4 Geographical analysis

As the map 2.1 shows it, Switzerland has a very specific topography. The north of the country is mainly flat, which enabled the development of a dense urban network dominated west to east by Geneva, Lausanne, Bern and Zurich. On the contrary, the south is a high mountainous region split between two major mountain chains, the Berner Alpen and the Glarner Alpen, with few cities of significant importance. This implies a lot of differences from a region to another as far as road geometry and road traffic are concerned. A geographical and topographical analysis of car crash risk in Switzerland is thus necessary. In this section, we will first study car crashes locations before comparing them to the road traffic, considered as an exposure variable (see 1.2.2), and analysing crashes "black-spots".

2.4.1 Where do car crashes occur mostly in Switzerland ?

The FEDRO bare car crashes locations database to which we had access for our study reports only the latitude and longitude of each bodily injury car crash location that occurred between 2011 and 2013. According to the privacy contract appending, we were not allowed to enrich this database with AXA Winterthur motor claims, nor enrich it to infer precisions about the circumstances of the crashes, such as the time slot or the severity (available only on an aggregate scale). We thus used here this database to better locate risky regions in the first place and then used it in a more proficient way while working with telematics data (see chapter 5).

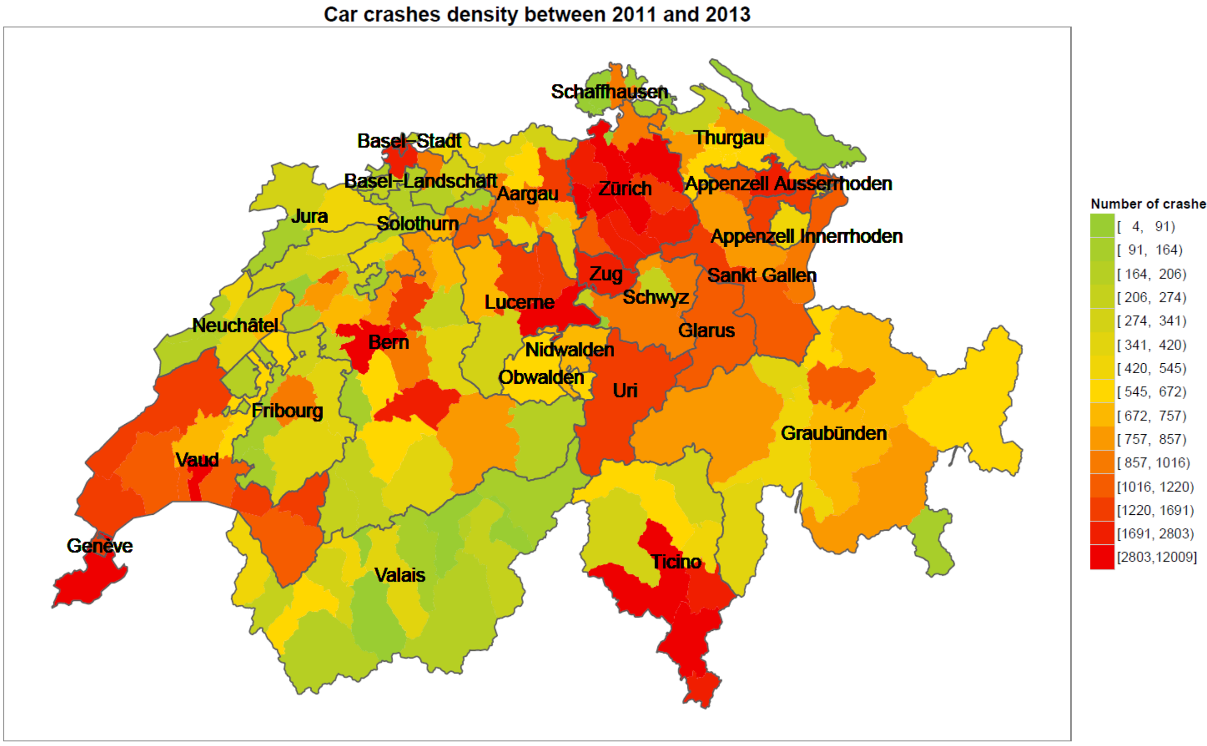


Figure 2.3: Density of car crashes per region

Figure 2.3 maps bodily injury crashes density per region. We notice first that Geneva and Zürich urban areas can be considered as extremely risky compared to the rest of the country, as their car crashes density amounts to more than 2 800 bodily injury car crashes. This has to be related to the number of inhabitants in these areas, which is relatively high (circa 372 000 inhabitants in Zürich, and 200 000 inhabitants in Geneva). Concerning the mountainous regions in the south of the country, we can spot three different type of regions. First Valais region, near Lausanne, is a non-risky region as its rashes density does not exceed circa 400 crashes. Second, Ticino region seems to be a very risky region as far as its car crashes density is the highest in the country. Moreover, this region is situated in the mountainous region linking directly Switzerland to the crowded region of Milan in Italy. This means that it is exposed to a lot of traffic, both tourist and freight, and has to deal with harsh road geometry and high declivity as well. Indeed from north to south of the region, within 150 kilometres, ground height decreases from 3000 meters high to only 200 meters near the Como lake. Finally, Graubünden region, located in the high mountain chain of Glarner Alpen, has a relatively median car crashes density (between circa 500 and 900 crashes). Linking it to the road network map 2.1, we notice that Graubünden road network is scattered, as well as in Valais region, which could explain this relatively low car crash density, despite harsh road geometry.

Nevertheless, as already mentioned, car crashes density analysis must be completed by a study of the traffic flow in each region. Indeed, increasing the number of vehicles travelling on a road leads to a raise in collision risk likelihood. How are traffic flow and car crashes density in each region related ?

2.4.2 Compare car crashes location to traffic exposure

In order to get a proxy of traffic flow in each region, we used data released each month by the the Swiss Automatic Road Traffic Counts (SARTC) network [12]. The Federal Roads Office owns a wide set of traffic counters all over the country. Those counters are located mostly on motorways and principal highways (see precise locations appended A.2). We averaged daily traffic over 2014 in each region, and deduced traffic for non-covered regions using a k-nearest neighbours algorithm. We thus got a map of the average daily traffic on principal highways over 2014 for each region (see figure 2.4). To be precise, we used 2014 daily traffic data to compare it to car crashes that occurred between 2011 and 2013, having checked that traffic flow trends were almost the same within the past three years.

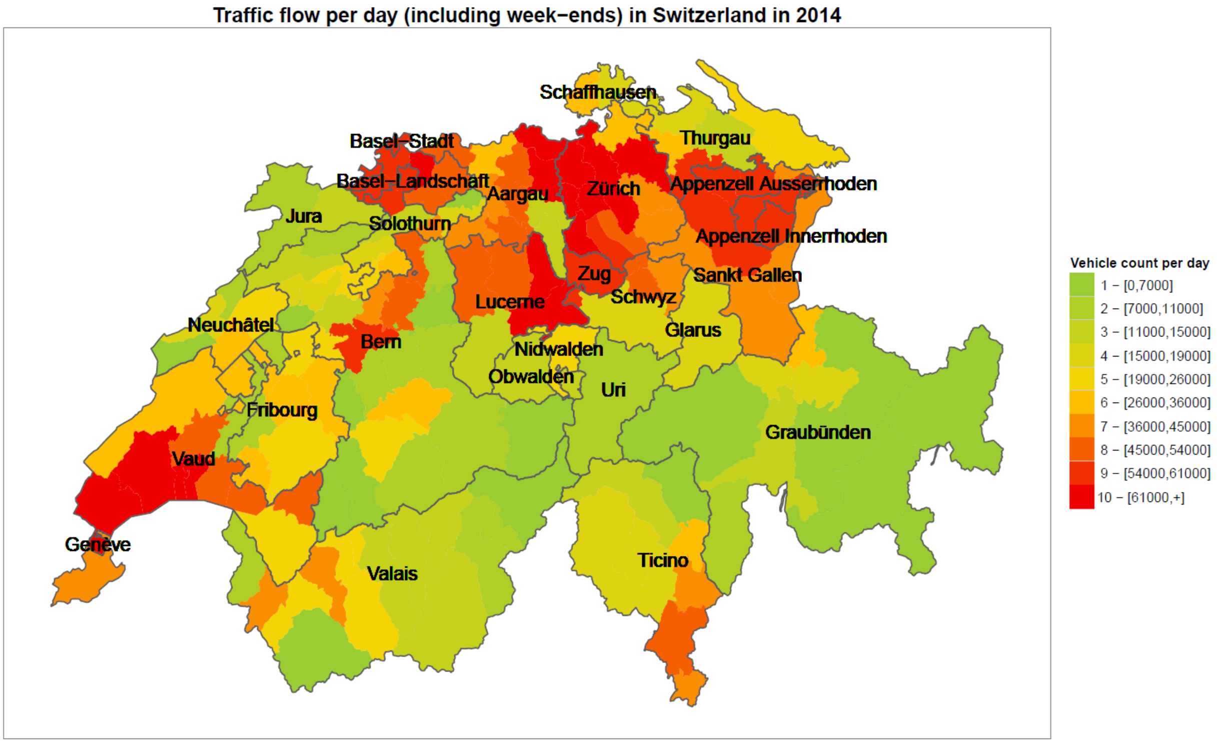


Figure 2.4: Density of traffic per region

As expected, we notice that traffic flow is far heavier in northern urban areas than in southern mountainous regions. A wide region around Zürich counts a heavy daily traffic flow, mostly coherent with the fact that Zürich is considered as the economic capital city of Switzerland. We observe also that the south of Ticino region has a relatively high traffic flow compared to the rest of the southern mountainous regions. This is

[12] Available online : http://www.astra.admin.ch/verkehrsdaten/00299/00301/index.html?lang=en

consistent with the fact that Ticino is a transition zone between Switzerland and Italy and has to cope with heavy freight and tourist traffic.

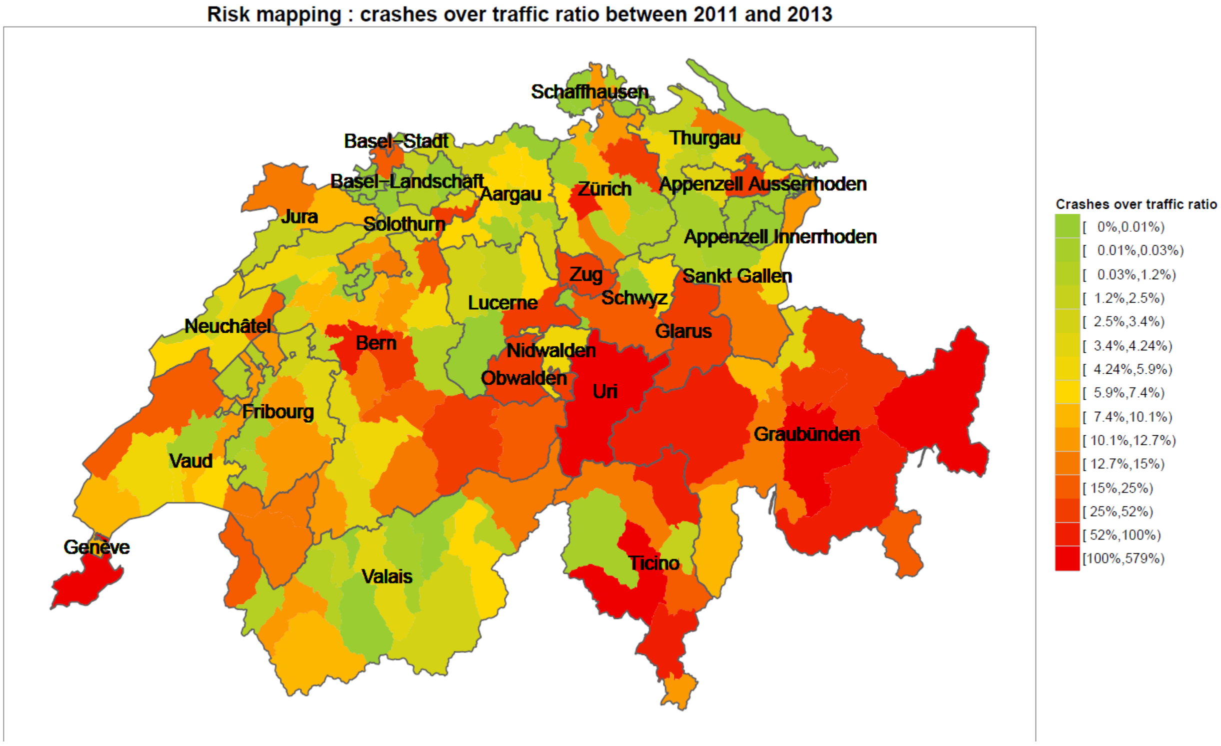


Figure 2.5: Ratio of crashes over traffic per region

Nevertheless, we notice that for some regions, the traffic map 2.4 is not in line with the car crashes density map 2.3. Indeed, Graubünden region, while having a median crashes count faces a very low daily traffic flow. This would mean that this region could be risky, understanding traffic as an exposure variable. To check these intuitions, we computed the ratio of car crashes count over daily traffic count for each region and obtained the map 2.5. It shows different risk profiles from one region to another :

- Geneva, Graubünden, and Ticino regions appear very risky, having a disproportionate car crash density compared to their daily traffic count.
- Zürich, having both heavy traffic and high car crash density, holds a proportionate car crash risk compared to its traffic exposition.

These geographical analysis favours partly our hypothesis on the inherent risk of roads due to their environment and geography. Indeed, for a same daily traffic flow (Valais and Graubünden regions for example), thus exposed *a priori* to the same car crash risk, thinking about traffic flow as an exposure variable, these two regions have strictly different risk profiles, certainly due to geographic and environment differences. We will test these hypotheses in the chapter focusing on telematics data (see part 5).

2.4.3 "Black spots" specificities

As mentioned above (see 2.2) the Swiss federal government issued a bill called the "Strassenverkehrsgesetz" (SVG) exclusively meant to globally improve road safety. This act provides a lot of measures, including a duty for the FEDRO to publish every year a report on the so-called "road safety black-points" (article 6a of the bill). This report is meant to analyse the Swiss road network from a road safety point of view, aiming at pointing out specific roads or locations that heavily need road maintenance. While improving road safety, this legal disposition will also enable the Swiss federal government to better target where their road maintenance budgets should be allocated. The first "black-spots" report [23] was issued in 2014 and focuses on bodily car crashes that occur between 2011 and 2013 all over the country [13]. We summarize here

[13] This report is based on the FEDRO bodily injury car crashes database presented in the first section of this chapter.

the main results of this study.

"Black-spots" research across the Swiss roads network is done according to tow major criteria : severity of the car crash (slight injury, severe injury, fatality) and road type (motorway, urban area road, countryside road). The analysis is done through the mapping of car crashes. Car crashes are associated to a "black-spot" location as soon as their characteristics (severity, road type) make their "norm" exceed a certain threshold. The "norm" used is a function of the number of severe crashes and slight injury crashes counted in a certain radius around a specific road, and overweights severe crashes (see appended specific rules A.1). roads are then ranked according to the total "norm" associated to their specific location.

This study points out 1 091 "black-spots", among which 91 are under the direct responsibility of the FEDRO. Among those "black-spots" a great majority is located at the junction of two roads of different scales : 61 are located at the junction of roads or streets of the secondary network, while 23 are located at motorway exits or ramps, the rest being located on common road sections. This insight will help us better understand the link between telematics data and claims locations. According to this report [23], 53% of the car crashes belonging to a "black-spot" are collision-type crashes, and 25% are linked to a control loss or to an off-road vehicle weaving. Studying more precisely the "black-spots" map published (see figure 2.6 [14]), we observe that they are mostly located around the major cities of Switzerland (Zürich specifically). Concerning the Graubünden region we mentioned above, we see that it counts only a few "black-spots" : this is certainly due to the fact that car crashes in this region are more uniformly spread on the roads than in more urbanised Swiss regions (see the appended map overlapping car crashes locations on traffic density A.1).

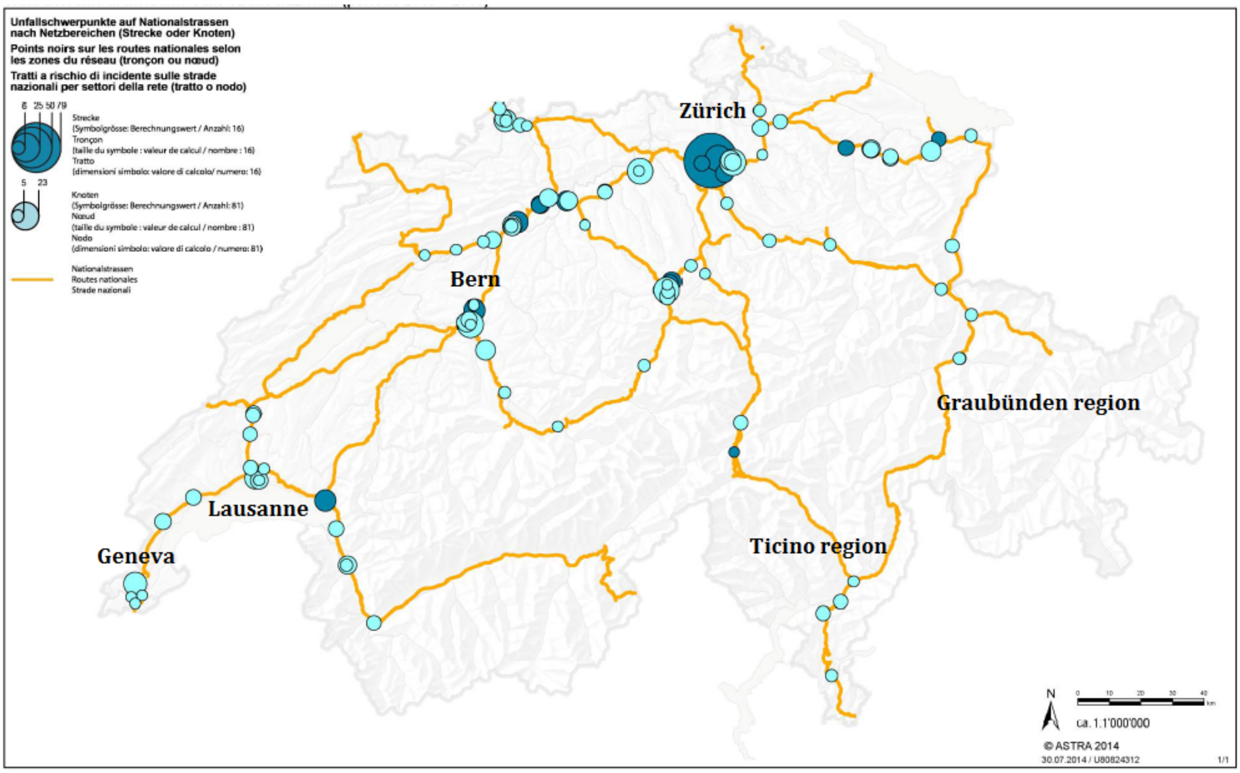

Unfallschwerpunkte auf Nationalstrassen nach Netzbereichen (Strecke oder Knoten)
Points noirs sur les routes nationales selon les zones du réseau (tronçon ou nœud)
Tratti a rischio di incidente sulle strade nazionali per settori della rete (tratto o nodo)
6 25 50 79
Strecke (Symbolgrösse: Berechnungswert / Anzahl: 16)
Tronçon (taille du symbole : valeur de calcul / nombre : 16)
Tratto (dimensioni simbolo: valore di calcolo/ numero: 16)
5 23
Knoten (Symbolgrösse: Berechnungswert / Anzahl: 81)
Nœud (taille du symbole : valeur de calcul / nombre : 81)
Nodo (dimensioni simbolo: valore di calcolo / numero: 81)
Nationalstrassen
Routes nationales
Strade nazionali
Zürich
Bern
Lausanne
Geneva
Ticino region
Graubünden region
N
0 10 20 30 40 km
ca. 1:1'000'000
© ASTRA 2014
30.07.2014 / U80824312
1/1

Figure 2.6: Map of "black-spots" locations reported by the FEDRO [23]

[14] The radius of each round on the map stands for the scale of the severity "norm" defined above.

2.4.4 Defining the geographical setting of our analysis

With knowledge of Switzerland geographic risk specificities, and because of technical matters related to big data treatment (see chapter 4), we decided to restrain our study to only four regions of Switzerland that appear to be of great interest for our purpose. Here is the scope we chose :

- Zürich region as it gathers urban areas features that are worth studying : both heavy traffic flow and high crashes density.
- Graubünden region as it is mostly mountainous and has a very specific risk profile, with a low traffic flow and a median car crashes density. We will thus be able to experience our hypothesis about road geometry and environment.
- Geneva region is also a urban area but with an environment very different from Zürich, as it is exposed to a lot of cross-boarder traffic on a restricted area located between a lake and the mountain.
- Ticino region seems to be risky as it has to cope with heavy freight and tourist traffic flow and is located in a mountainous area.

PART 3

A temporal analysis of daily claims count

Insurance business activity is characterized by a so-called "inverted production cycle" : insurers get paid for a service they would *potentially* deliver in the future to their customers. This service consists in compensating insured customers for losses they were hedged against thanks to the premium they pay. Therefore, insurers must set aside enough provisions in order to keep being solvent while compensating their customers when claims occur. Consequently, asset-liability management is done mainly according to risk quantifications and predictions. Indeed, a better risk forecast entails a more accurate provisioning. Moreover, by performing those risk analyses, insurers participate to safety enhancement. As far as retail insurance is concerned, risks specification is generally carried out by modelling claims frequency and average costs assuming conditional independence between them. For the moment, risk analyses are generally performed *yearly*. In this chapter, we seek improvement in claims count modelling by carrying out *daily* analyses enriched with external temporal data : we will model AXA-Winterthur daily claims count using time series analysis and external weather and traffic data.

3.1 Data introduction

AXA Winterthur motor insurance portfolio is composed of nearly one million customers living all across Switzerland, among which more than a half subscribed to a collision coverage. We focus here on these latter customers and on *collision* claims involving personal vehicles only recorded by AXA-Winterthur between 2008 and 2014. Commercial vehicle claims were excluded from this database in order to better capture retail motor risk only. For each claim recorded in the database, its occurrence date only was available. We thus counted the number of claims AXA had to cope with for each day and got a daily claims count over the period 2008 to 2014.

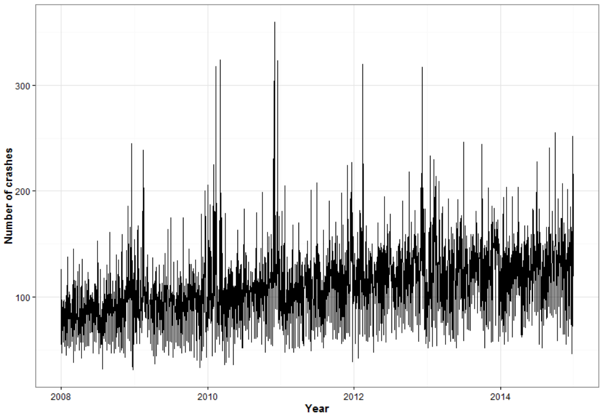


Figure 3.1: Daily claims count time series

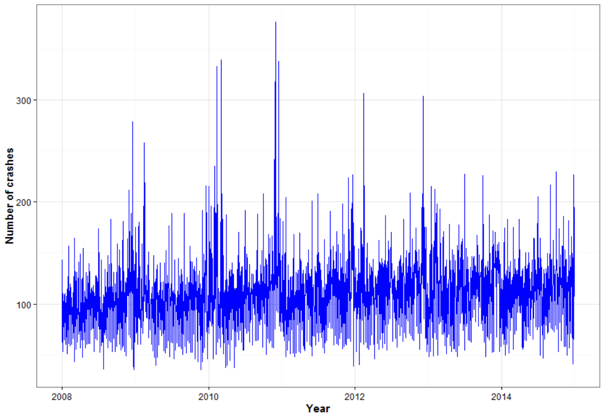
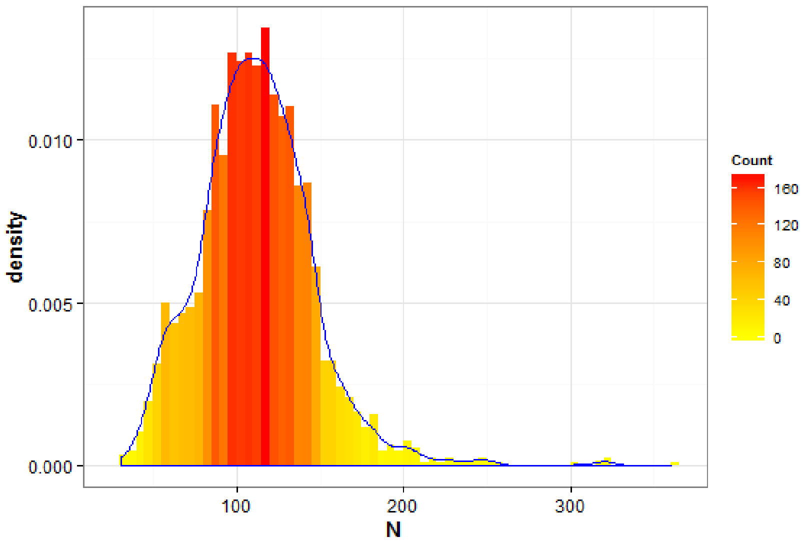


Figure 3.2: Corrected daily claims count time series

Figure 3.1 shows the time series derived from the claims recording. At first sight, we see an upward trend from 2008 to 2014. It can be interpreted in two different ways : either insured people were more risky in recent years, or it is due to a portfolio size effect. In order to check it, we divided daily claims counts by the yearly portfolio size associated and got figure 3.2 [1]. We hence observe that almost all the trend previously spotted was removed, as the global trend is now relatively flat. AXA-Winterthur's retail motor portfolio

[1] In order to keep an understandable scale, we multiplied claims count over portfolio size ratios by portfolio's size mean over the period 2008 to 2014.

size raised indeed by nearly 26.4% between 2008 and 2014, partly because of AXA and Winterthur merging in 2007 that impacted retail business lines in 2009. We will take this scale effect into account in our further count data modelling (see 3.3.3).
Then, looking closer at the chart, we note that peaks are almost steadily spaced, especially huge ones. It implies that our time series may be impacted by some seasonal effects. We will test this hypothesis using time series theory (see 3.2).
Finally, this dataset is composed of *count data* (i.e. *positive integer* data) as figure 3.3 shows it, which implies statistical modelling based on specific probability distributions. Among them are the Poisson and the Negative Binomial distributions, which will be closer examined section 3.3.



	Claims count
Minimum	31
Median	110
Mean	111.9
Maximum	360
Variance	1 204.03

Figure 3.3: Distribution of actual daily claims count

3.2 Time series analysis

We examine in this section daily claims count from a time series point of view. From the dates associated to each claim count, we derived several time features such as the month, the day of the week, the season, etc. in order to better portray seasonal effects spotted. A graphical calendar depiction of the series will be explored before introducing more technical tools for time series analysis. To be precise, our goal here is to extract meaningful time-based features in the daily claims count time series, that will be plugged in an econometric model afterwards, and not to fit an exhaustive time series model.

3.2.1 Calendar view

Figure 3.4 shows a calendar heat map view of AXA-Winterthur daily corrected [2] claims count. This type of graphical representation is very useful to get some insights on the possible seasonal effects at stake. The darker the cell is, the more claims happened on that day. Several peculiarities are worth mentioning :

- Claims count is much lower on Sundays and much higher on Mondays than on any other day of the week, whatever the month or the year concerned.
- Winter months (November, December, January and February) record higher claims count, whatever the year. It may be put down to harsher weather conditions : we will test this hypothesis in our count data modelling (see 3.3).
- 2010 seems to have been the riskiest year within the period, as the highest claims counts occurred that year (see the brownest cells of the calendar heat map).

[2] In order to better spot pure timely seasonal effects in this graphical representation of our dataset, we plotted here claims count time series corrected from the portfolio size effect computed as explained above (see 3.1).

- On the first day of each month, claims count records peaks. Nevertheless, there seems to be no valid and coherent reason to explain this phenomenon. We assumed that it could be due to a database bias : when crash occurrence date is misreported or not certainly known, a default setting may set it to the first day of the month on which it happened.
- Public holidays must also be carefully examined, as claims count on the days preceding a Swiss public holiday seem to be higher than on other days.

We thus expect to capture at least a weekly seasonality in our time series while using time series theory.

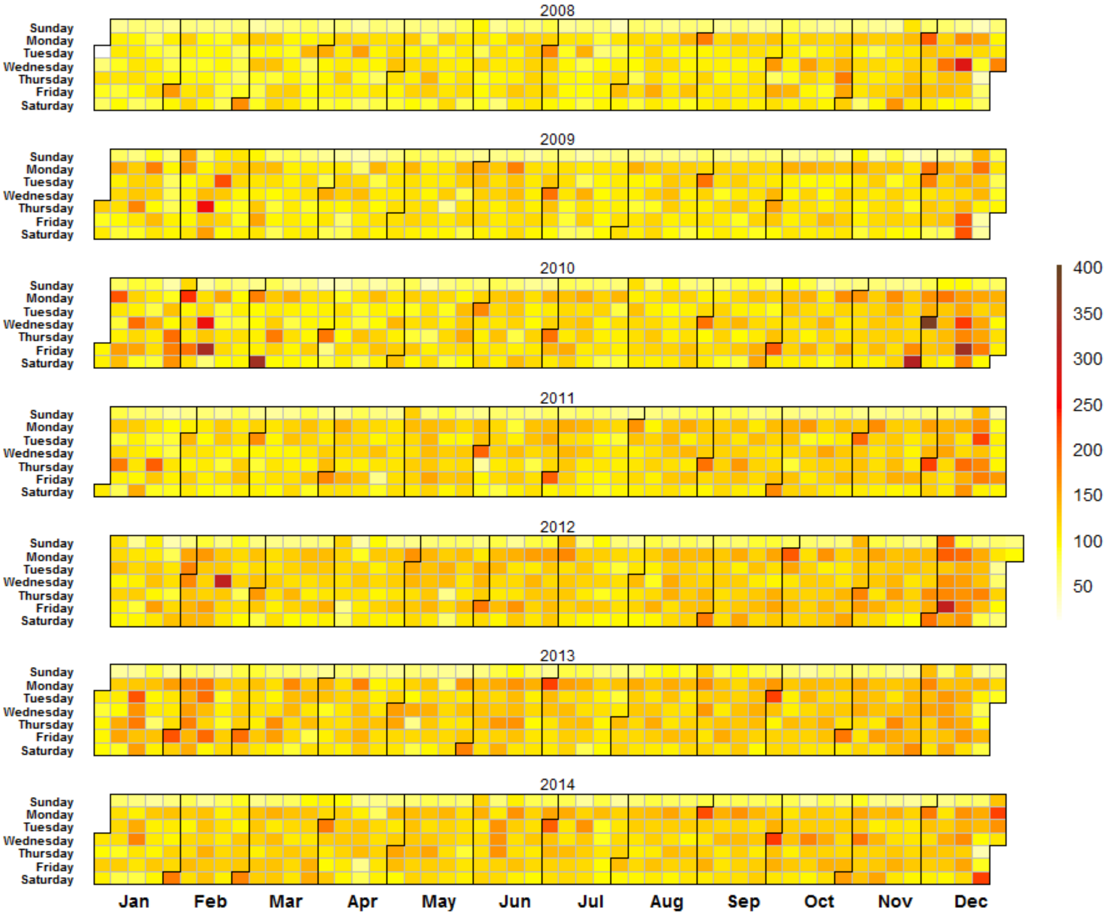


Figure 3.4: Calendar heat map view of corrected daily claims count time series

As mentioned above (see 1.2.2), traffic count must be taken into account as an exposure variable, as soon as we speak about car crashes. Indeed, aggregating individual risks entails heavier global risk [3]. It is even more valid when looking at *collision* crashes, as they involve at least two vehicles : a heavier traffic flow entails reasonably greater collision risk. In order to get a proxy of daily traffic flow over the period 2008 to 2014, we used monthly data from the Swiss Automatic Road Traffic Counts (SARTC) network [4]. As the claims counts examined are not geo-located, we only use the temporal dimension of the traffic database in order to match it with the claims count one. We thus averaged, for each month over the period, traffic counts of all the stations across the country, and got a monthly time series of traffic count in Switzerland. Figure 3.5 shows a calendar heat map of the monthly traffic count computed. As expected, we notice that traffic flow is far much lower on Sundays than on other days of the week whatever the year or the month. Besides, we remark that traffic flow was heavier in 2010 than in the other years of the period watched : it may partly explain 2010 increased claims rate. These graphical insights must now be corroborated by more technical further analysis, using count data models especially (see 3.3).

[3] From a statistical point of view, if we model each individual car crash risk using a Poisson distribution $\mathcal{P}(\lambda)$, assuming n independent risks, global risk distribution will result in a Poisson distribution $\mathcal{P}(n \times \lambda)$, n being thus an exposure variable. Independence between λ and n still remains at stake.

[4] For further information about this database, see 2.4.2.

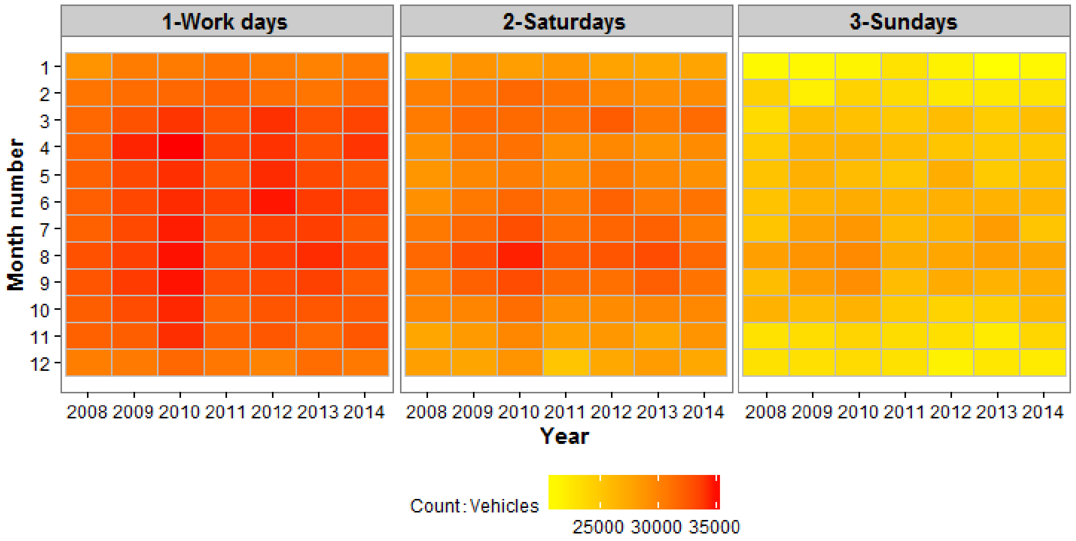


Figure 3.5: Calendar heat map view of monthly traffic flow

3.2.2 Looking for seasonality features using time series theory

The following analyses were performed on the *raw* daily claims count time series (see figure 3.1). We first examine autocorrelations and unit root tests, then study how the series can be decomposed between a trend feature and a seasonal one, and finally try to explain in plain words to which time periods the seasonal features extracted can refer.

Unit root tests

Figure 3.6 shows the autocorrelogram and the partial autocorrelogram derived from the series. The autocorrelogram computed shows a clear seasonality in the daily claims count time series, as it records an obvious peak every seven days. It thus corroborates the intuition we had looking at the calendar view (namely that claims count behaved differently according to the day of the week). Furthermore, it may give a strong signal of non-stationarity of the series.

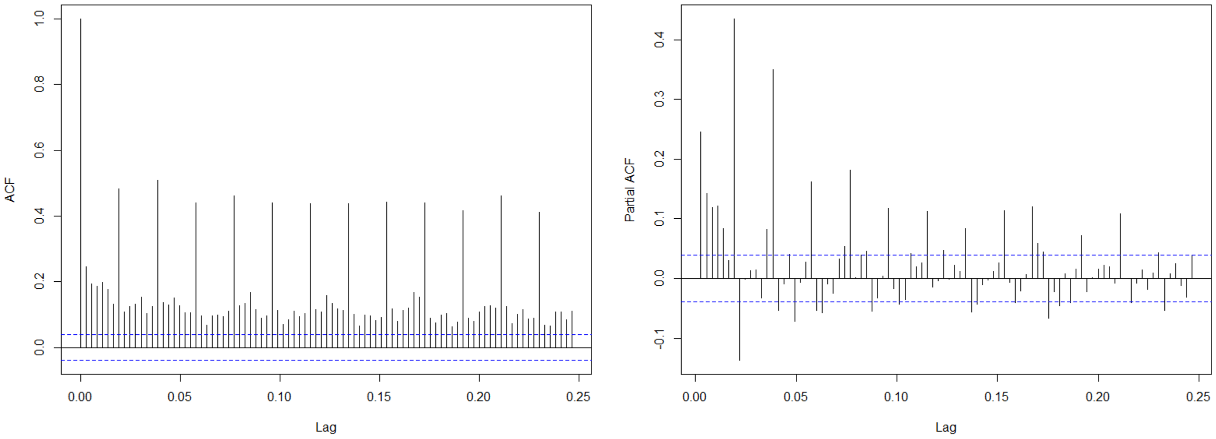


Figure 3.6: Autocorrelogram (left) and partial autocorrelogram (right) of daily claims count time series

In order to check this hypothesis, we performed the standard Augmented Dickey-Fuller (ADF) unit root test[5] [6] which results are displayed in table 3.1 [7].

[5] To confirm ADF test results, we also carried out a Schmidt-Philips test, which results are available appendix B.1.

[6] For further information about Augmented Dickey-Fuller and Schmidt-Philips tests, see [10] and Wikipedia pages associated.

[7] Codes : '***' states for a 0.1 % significance, '**' for a 1% significance, '*' for a 5% significance and '.' for a 10% significance

Augmented Dickey-Fuller test (ADF) As we expect to find a significant trend in the time series (see 3.1), we use here the "trend" version of the Augmented Dickey-Fuller test. This means that, calling the series y_t, the null hypothesis $H_0 : \alpha = \beta = 0, \rho = 1$ is tested, assuming that :

$$\Delta y_t = \alpha + \beta \times t + (\rho - 1) \times y_{t-1} + \sum_{i=1}^{p-1} \alpha_i \times \Delta y_{t-i} + \epsilon_t \qquad (3.1)$$

with α being the drift, β being the linear trend coefficient, p a chosen lag number (we fixed it at 7, in order to capture the weekly seasonality) and Δy_{t-i} the i^{th} order differentiated time series y_t. In plain text, if H_0 is not rejected, it means that our time series is non-stationary and encompasses neither drift nor trend. More precisely, the ADF test computes three test-statistics : τ_3 tests only the presence of a unique root ($\rho = 1$) i.e. if the series is integrated of order 1, Φ_2 tests both the presence of a unique root and a drift combined to linear temporal trend ($\alpha = \beta = 0, \rho = 1$), and Φ_3 tests both the presence of a unique root and a drift only ($\alpha = 0, \rho = 1$).

Coefficient	**Estimate**	**p-value**	**Significance**
α	52.43	< 2e-16	***
β	0.01	< 2e-16	***
ρ	0.40	< 2e-16	***
α_1	-0.22	1.71e-07	***
α_2	-0.22	4.01e-08	***
α_3	-0.22	4.66e-09	***
α_4	-0.18	1.06e-07	***
α_5	-0.17	6.19e-09	***
α_6	-0.25	< 2e-16	***
α_7	0.17	< 2e-16	***

	Test-statistic	**1% critical value**
τ_3	-13.94	-3.96
Φ_2	64.80	6.09
Φ_3	97.17	8.27

Table 3.1: Augmented Dickey-Fuller unit root test results

Looking at the results, both tests reject the null hypothesis that y_t is integrated of order 1 [8]. It means that we need to introduce an auto-regressive parameter in our further econometric model, but instead of including it as an offset, we will estimate the coefficient associated. At the same time, the ADF test performed rejects the null hypothesis $H_0 : \alpha = \beta = 0, \rho = 1$ (i.e. that there is neither drift nor linear temporal trend but a unit root), if we look at ϕ_2 test-statistic value. Furthermore, in both tests, α and β estimates are significant, showing the hypothesis according to which there is an upward trend in the time series is validated. Besides, β estimate is circa 1% in ADF test and 2% in Schmidt-Phillips test (t being years) : these figures are coherent with the actual level of AXA's annual portfolio size growth over the period 2008 to 2014, which amounts on average to 4%. It confirms partly that the time series upward trend must be assigned to portfolio growth over the years.

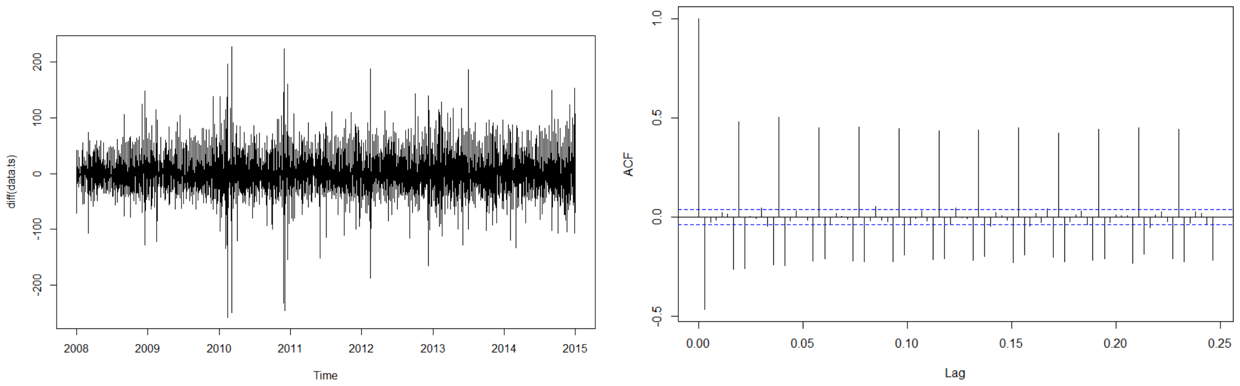


Figure 3.7: Plot and autocorrelogram (right) of differentiated daily claims count time series

[8] τ_3 test-statistic is taken into account to draw this conclusion.

Then, lag coefficients estimated by the ADF test (α_1 to α_7) are all significant, which confirms that a weekly seasonality is worth spotting. Indeed, looking at the differentiated series chart (see figure 3.7) and its associated autocorrelogram, a weekly seasonal feature is obviously driving the series. Moreover, sixth and eight autocorrelation lags are negative, explaining partly why there is a sign inversion between α_1 to α_6 and α_7.

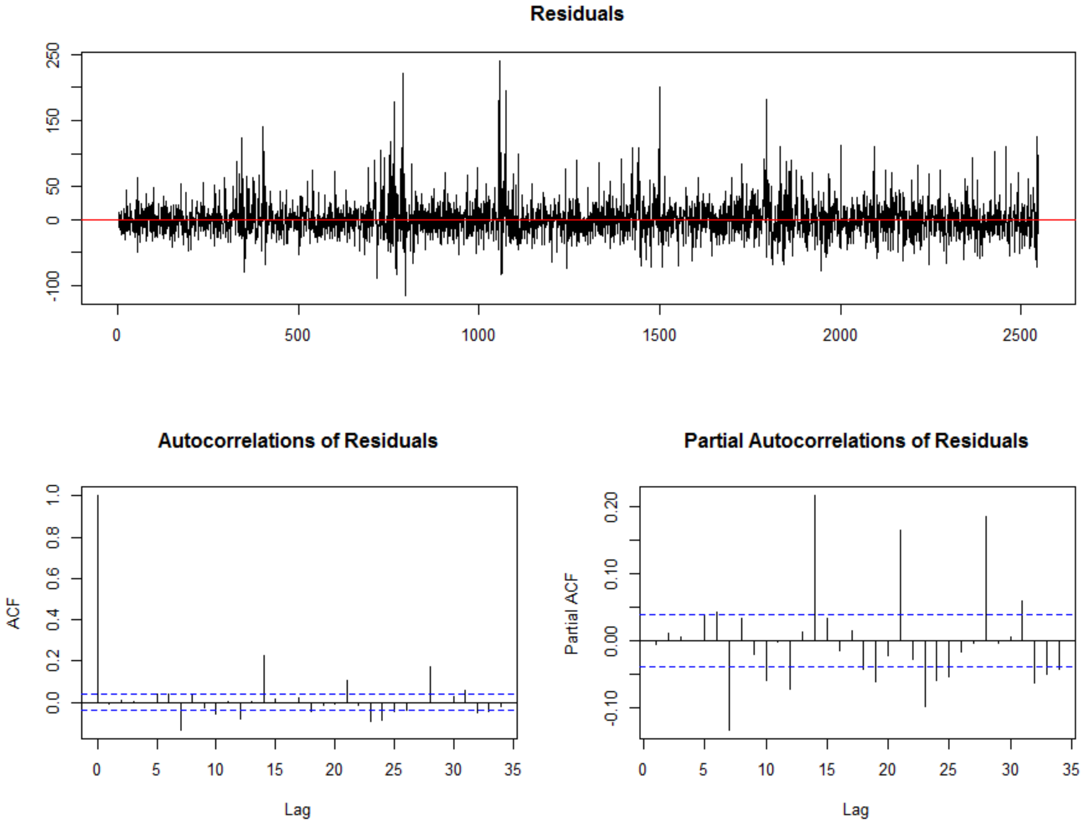


Figure 3.8: ADF tests residual analysis

Finally, figure 3.8 shows residuals extracted from the ADF regression model and the autocorrelograms associated. Looking closer to the residuals autocorrelogram, we see that almost all the seasonal autocorrelation peaks were removed performing the ADF regression estimation. Nevertheless, we can still spot peaks in the residual series chart that seem to be closely gathered : it means that the model potentially faces strong heteroscedasticity [9]. Besides, the associated partial autocorrelogram is far from being flat, advocating for moving average features (i.e. lags on error terms) inclusion in our model. However, as our goal is only to point out meaningful temporal features, we decided neither to fit an heteroscedastic model nor to integrate moving average features, that would have been too intricate.

Trend-seasonal decomposition

In order to better catch how drift, trend and seasonal elements are nested, we used a classical trend-seasonal decomposition by moving average. The additive model used writes :

$$y_t = T_t + S_t + e_t$$

T_t being the trend, S_t being the seasonal scheme and e_t being the error term. Figure 3.9 displays each term of the trend-seasonal decomposition : it confirms the presence of a trend and strong seasonal features that

[9] We indeed observe in the residuals chart that periods of low variance (that is low amplitude) are followed by periods of high variance. For instance compare the series amplitude from the first to the 300^{th} with the series amplitude reached between the 700^{th} and 800^{th} observations.

were highlighted by the ADF test parameters estimation. This decomposition is computed in three steps that summarize briefly to :

1 . Trend component extraction using a classical moving average with an equally-weighted symmetric window.
2. Seasonal component computation on the remaining signal (raw time series from which the trend component extracted was removed) by averaging and centering it for each time unit over all period.
3. Error component extraction by removing both the trend and the seasonal components previously computed from the raw time series.

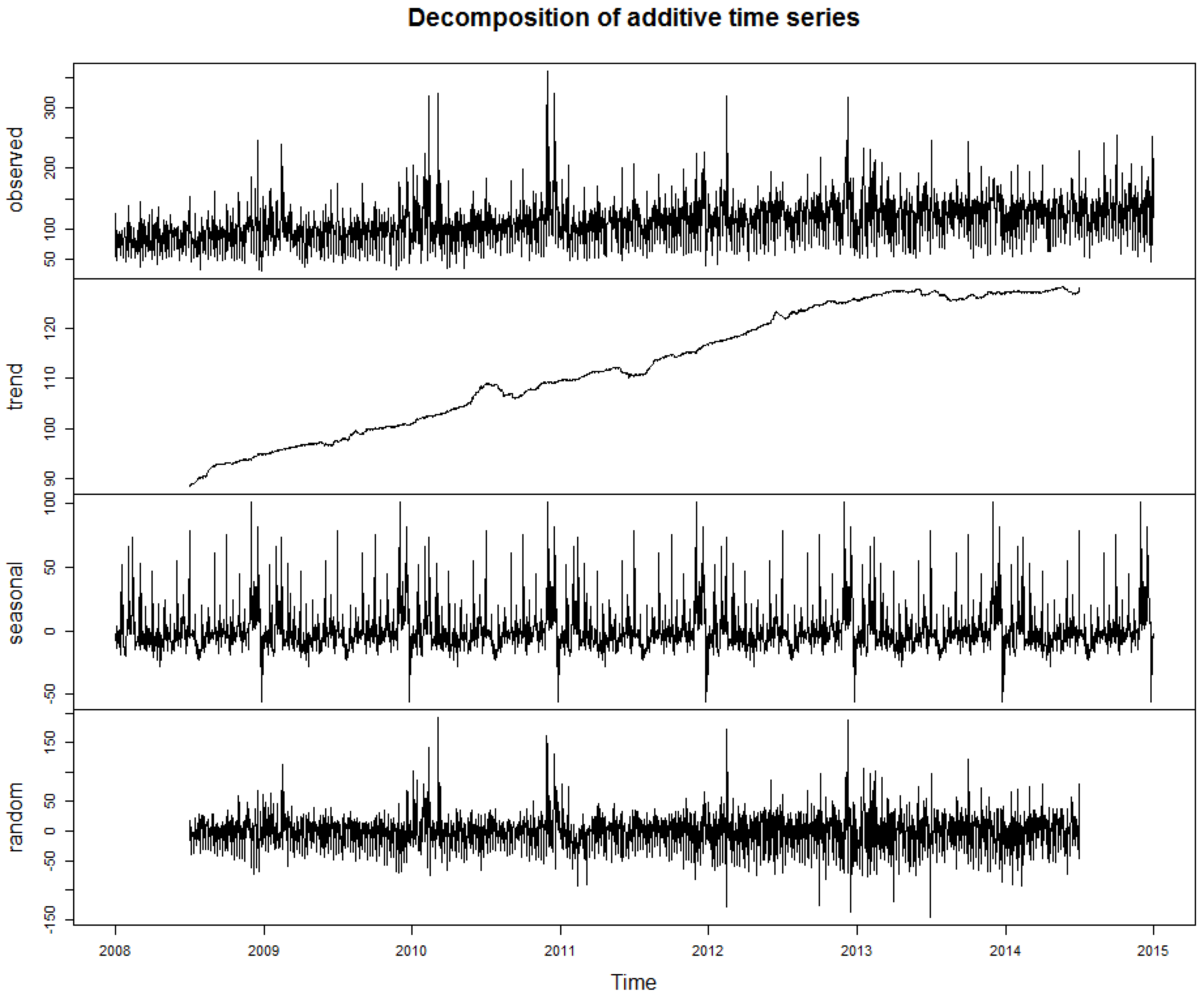


Figure 3.9: Trend-seasonal decomposition of daily claims count time series

Assessing seasonality effects

Now our time series is decomposed in a trend component and a seasonal component, we want to explain in plain words to which phenomena they may be due. We already explained potential reasons that lead to the trend component extracted (see 3.1). Examining again the calendar view figure 3.4 and the insights derived from it, we can point out three possible time units that may affect daily claims count :

- Year : As the trend extracted is highly significant in the time series depiction and linked to *yearly* portfolio size growth, it may be significant.
- Month or season : As winter months seem to be more risky than others whatever the year, it may be included in our model as a seasonality parameter.
- Day : Since traffic exposure depends on days (week-days being over-exposed, and Sundays being under-exposed), a "day" feature must be added to our further claims count model.

- Holidays : As driving habits can be strongly influenced by holidays (longer trips, changing driving hours ...), it is essential to take this parameter into account.

Average claims count were thus computed for each time period discussed above. AXA daily claims count time series was then centered round those averages. In order to check if all seasonality effects are removed using those four features, we plotted the autocorrelogram and partial autocorrelogram associated (see figure 3.10). Seasonality was obviously sharply reduced using this approximate seasonality adjustment, as no more regularly spaced autocorrelation peaks are noticing. There still remain a relatively high autocorrelation at first lag. It means that claims count recorded on the day before should be taken into account in the model (feature that was already enlightened by ADF test estimation).

Besides, the associated partial autocorrelogram advocates for moving average features addition. Nevertheless, as already mentioned, moving average features are far more difficult to interpret than autoregressive parameters. That is why we did not fit a SARIMA model, nor tried to develop a more complex time series model that would have better fitted our dataset.

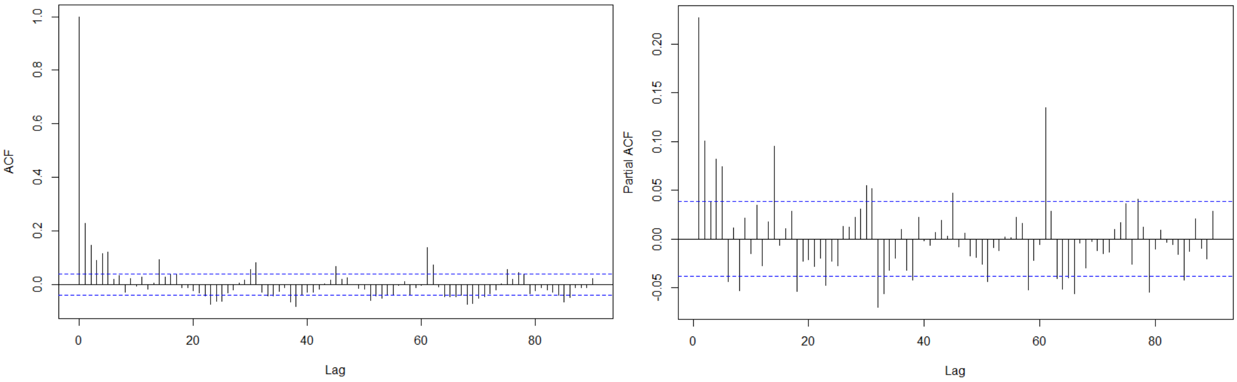


Figure 3.10: Autocorrelogram (left) and partial autocorrelogram (right) of seasonally adjusted daily claims count series

To conclude, this preliminary time series work pointed out some specific temporal and external features that need to be included in our further claims count modelling : year, season, day of the week, claims count on the previous day, public holidays and exposure parameters such as portfolio size and traffic count.

3.3 Count data models

As already highlighted, our dataset is composed of *count* data, which entails specific modelling that will be carefully examined in this section. External weather data is first introduced, then the pros and cons of Poisson and Negative Binomial count data models are discussed considering the dataset at stake.

3.3.1 Weather data treatment

The National Oceanic and Atmospheric Administration (NOAA) is an American scientific federal agency working on global atmospheric and weather conditions on behalf of the World Meteorological Organization (WMO). It owns weather stations all around the world and especially in Switzerland and displays freely weather data recorded by these stations [10]. The Swiss network of NOAA stations is composed of circa 230 stations [11], which record *hourly* parameters such as air temperature, weather depiction, visibility distance and wind speed [12].

[10] Weather data from the NOAA are freely available on the following website : http://www.ncdc.noaa.gov/data-access

[11] A map of NOAA weather stations locations is appended B.1.

[12] Actually, more specific variables such as precipitations amounts or cloud density are recorded by a few stations, but these data fields are almost always not available. That is why we focus here on the standard variables mentioned above.

These weather data are thus far more detailed than the daily claims count data studied. Moreover, contrary to claims data, weather data is geo-located. Thus, we had to summarize *hourly* data into a *daily* format and average it at a country level, in order to accurately match both datasets. More precisely, parameters values were first averaged over a day for each station, and then these averaged values per station were gathered and averaged over all the stations of the country. The summarizing values chosen for each parameter are :

- Wind speed : daily average
- Air temperature : daily median
- Visibility distance : daily average
- Weather depiction : most recurrent weather depiction over a day (verbal depictions were classified as : "cloudy", "fog", "rain", "snow", "thunderstorm" and "nothing significant" meaning roughly fine weather)

3.3.2 Modelling daily claims count using count data econometrics

Our target is to model AXA-Winterthur daily claims count recorded between 2008 and 2014. Let's call this variable Y, y_t being the observed realization of Y on day t. In order to explain this target variable, we use the following range X of time-based features (derived from time series studies, see 3.2) and previously introduced weather features [13] :

Covariate	**Type**	**Meaning**	**Unit**	**Baseline modality**
Temporal variables				
Year	Quant.	Observation year	-	-
Season	Qual.	Observation season	-	Spring
Day	Qual.	Observation day (Monday, Tuesday ...)	-	Wednesday
Holidays	Dummy	Equals 1 if the observation day was a public holiday	-	-
Autoregressor	Quant.	Claims count on the day before the observation	Count	-
Weather variables				
Air temperature	Quant.	Averaged air temperature recorded by Swiss NOAA weather stations on the observation date	Celsius degrees	-
Wind speed	Quant.	Averaged wind speed recorded by Swiss NOAA weather stations on the observation date	Meters per second	-
Visibility distance	Quant.	Averaged visibility distance recorded by Swiss NOAA weather stations on the observation date	Kilometres	-
Weather depiction	Qual.	Most recurrent weather depiction recorded by Swiss NOAA weather stations on the observation date	-	Nothing significant (i.e. fine weather)
Exposure variables				
Traffic count	Quant.	Average number of vehicles registered on Swiss highways by SARTC traffic counters on the type of day (business day, Saturday or Sunday) and month of the observation date	Bunches of 5000 vehicles	-
Portfolio size	Quant.	Number of customers in AXA's retail motor insurance portfolio during the observation year	-	-

Table 3.2: Covariates of fitted claims count models

[13] "Qual." states for "qualitative" and "Quant." states for "quantitative". Baseline modalities were chosen according either to modality meaning or to modality highest recurrence.

Graphical insights

Before trying to fit any model, univariate charts were plotted in order to get some insights about how the covariates chosen can affect daily claims count. Some examples of these charts are available figures 3.11, 3.12 and 3.13. For instance, figure 3.11 obviously show that higher claims count are recorded on business days than on week-ends (88% more claims are recorded on Mondays than on Sundays on average). Figures 3.12 and 3.13 highlight the potential downward effect of public holidays and upward effect of snowy weather on daily claims count (+51% on average). These effects are expected to be caught by count data models.

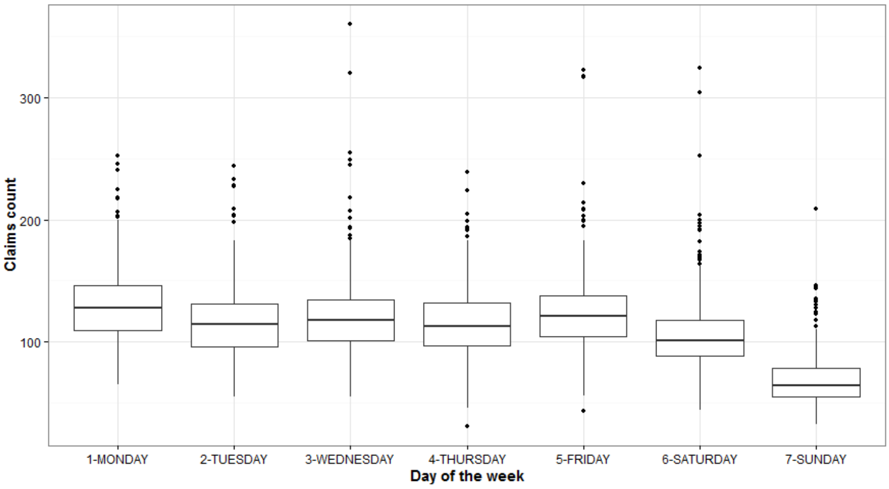


Figure 3.11: Claims count dispersion according to the type of day

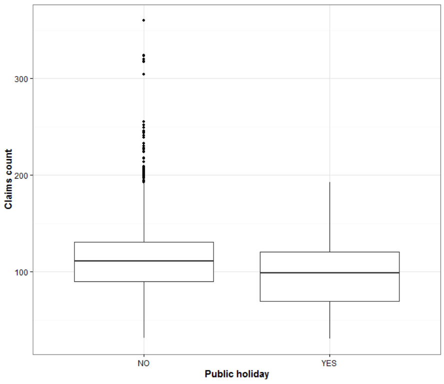


Figure 3.12: Claims count dispersion on public holidays

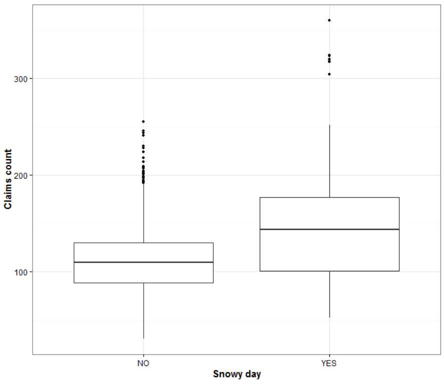


Figure 3.13: Claims count dispersion on snowy days

Which statistical distribution would be more appropriate to model this count dataset ?

As already mentioned, claims data is *integer positive*. A simple linear regression model is thus totally inappropriate to fit it. Generalized Linear Models (GLM) based on exponential-family distributions were thus considered. Among those distributions, we decided to test both Poisson and Negative Binomial distribution (both using logarithm link), as their supports are both $\mathbb{N}$.

One of the main differences between Poisson and Negative Binomial distributions lays in data dispersion. A Poisson distribution is suitable for fitting equally dispersed data, as if $X \sim P(\lambda)$, $\mathbb{E}(X) = \mathbb{V}(X) = \lambda$. Conversely, a Negative Binomial distribution is more appropriate to fit over-dispersed data [14] because if $X \sim NB(r, p)$ [15], then $\mathbb{E}(X) = pr/(1-p) < \mathbb{V}(X) = pr/(1-p)^2$. Negative Binomial regression can be considered as a generalization of Poisson regression since it has the same structure as Poisson regression but includes an extra parameter to model over dispersion [16].

[14] A data series is said to be over-dispersed when $\mathbb{V}(Y|X) = \gamma\mathbb{E}(Y|X)$ with $\gamma \gg 0$.

[15] $r > 0$ is the number of failures until the experiment is stopped and $p \in [0, 1]$ is the success probability in each experiment

[16] Called σ in estimation results table 3.5.

To be more accurate, "over-dispersion" in a GLM framework refers to an "over-dispersed" *conditional* distribution. Indeed, a GLM seeks to fit $\mathbb{E}(Y|X)$. The model writes, g being the link function and $\mathbb{P}_\theta$ being the exponential family distribution chosen (Poisson or Negative Binomial in this case) :

$$Y|X \sim \mathbb{P}_\theta \qquad \text{with} \qquad g(\mathbb{E}(Y|X)) = \beta' X \tag{3.2}$$

Moreover, *conditional* $\mathbb{P}(Y|X)$ and *non-conditional* $\mathbb{P}(Y)$ distributions are related by the Bayes formula : $\mathbb{P}(Y) = \mathbb{P}(Y|X) \times \mathbb{P}(X)$. Thus looking for over-dispersion in the raw daily claims count empirical distribution [17] (i.e. looking at the *non-conditional* distribution) is misleading and not helpful to choose between the Poisson and Negative Binomial distribution. Furthermore, $\mathbb{P}(X)$ being unknown, we cannot favour *a priori* a distribution over the other. This justifies why we tested both generalized linear models.

Nevertheless, looking closer at the data, we presume our target is *conditionally* strongly over-dispersed, which means the Negative Binomial model fit should be better than the Poisson one. In order to empirically check it, the target variable Y was clustered with respect to covariates X values to approximate $Y|X$ distribution . Conditional mean, variance and dispersion (variance over mean ratio) of daily claims count were computed within each group of weather depiction and day. Table 3.3 obviously shows that variance is much higher than mean within each group, advocating for *conditional* over-dispersion [18]. We must however acknowledge that Negative Binomial model estimates only one dispersion parameter over the whole conditional distribution $Y|X$, meaning that the dispersion parameter should be almost the same within each group. We observe that it is not the case for the segmentations tested (table 3.3) possibly due to heteroscedasticity effects. Model fitting will confirm or overturn these presumptions.

Weather depiction	Mean	Variance	Dispersion
Fine weather	111.62	933.41	8.33
Cloudy	108.20	164.70	1.52
Rain	110.45	978.94	8.87
Fog	108.72	971.29	8.93
Thunderstorm	136.44	1 398.78	10.28
Snow	149.49	3 911.65	26.17

Day	Mean	Variance	Dispersion
Monday	129.89	865.61	6.65
Tuesday	116.27	769.03	6.63
Wednesday	120.73	1 053.46	8.73
Thursday	116.59	873.59	7.49
Friday	123.37	1 039.36	8.42
Saturday	107.44	948.21	8.86
Sunday	68.83	446.29	6.46

Table 3.3: Mean and variance of daily claims count per weather depiction and type of day

Dealing with exposure variables

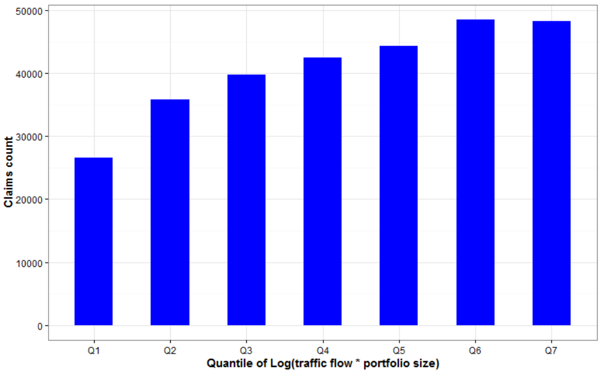


Figure 3.14: Bar plot of claims count sum within each quantile of the offset parameter

Traffic exposure and portfolio size were both spotted as exposure variables in the previous analyses (see 3.2). In order not to bias further econometric analyses, it is thus essential to normalize daily claims count series with respect to those variables. As the GLM models further tested (Poisson and Negative Binomial

[17] Looking at figure 3.3 page 36, daily claims count is obviously strongly *non-conditionally* over-dispersed : its variance over mean ratio (also called "dispersion") amounts to 10.75.

[18] Our segmentation check is however certainly not enough accurate to completely detect *conditional* over-dispersion.

ones) both rely on a logarithmic link [19], target normalization boils down to :

$$\log\left(\frac{\mathbb{E}(Y|X)}{\text{Traffic} \times \text{Portfolio size}}\right) = \beta' X \qquad \Leftrightarrow \qquad \mathbb{E}(Y|X) = \exp\left(\log(\text{Traffic} \times \text{Portfolio size}) + \beta' X\right)$$

That is why a $\log(\text{Traffic} \times \text{Portfolio size})$ *offset* was included in the formula of both GLM models fitted. Figure 3.14 illustrates the relation between daily claims count (our target) and the offset chosen : a clear upward linear trend can be spotted, which confirms our intuitions [20]. It implies that when interpreting the β coefficients, we look at the impacts of covariates on the proportion of daily claims count per driver.

3.3.3 Results

Both a Poisson and a Negative Binomial regression models were fitted on the data (see 3.3.2) : tables 3.4 and 3.5 gather marginal effects estimation results[21]. Going through the **Poisson model** results (table 3.4), we observe that almost all variables included in the model are significant at a 0.1% level. Among those significant effects, some are worth explaining :

- Even if the size of retail motor AXA portfolio was taken into account as an offset parameter, daily collision claims count per driver raises by 1.15% per year.
- Collision car crashes count per driver inflate by 10.10% in Winter compared to Spring season.
- Claims count per driver is lessened by 25.14% on Sundays compared to Wednesday record.
- Collision crashes count per driver is reduced by 18.49% on public holidays.
- Counting one more claim on the day before leads to 0.13% more claims on the day under scrutiny.
- If the visibility distance is enhanced by 1 km, claims count per driver is dialled down by 8.56%.
- Crashes count per driver increases by more than 25% on snowy days.

Now looking at the **Negative Binomial** model results (table 3.5), all coefficients estimated are remarkably almost the same as in the Poisson model, turning out to the same interpretations, but less coefficients are significant (wind speed and air temperature especially are not significant anymore). The dispersion parameter σ estimate amounts to nearly 38.40 with a high significance (low standard error).

Models comparison

In order to compare Poisson and Negative binomial models performance, we studied standard AIC criteria, carried out a likelihood ratio test and used deviance criteria.

→ Roughly analysing **AIC criteria**, Negative Binomial model seems to be more appropriate than the Poisson one, as Negative Binomial AIC is far much lower than Poisson model AIC. But this conclusion needs to be confirmed by other criteria.

→ Computing **likelihood ratio** test-statistic [22], we get : $D = 4354.213$. The null hypothesis is thus rejected [23], strongly suggesting that the negative binomial model, that estimates a dispersion parameter, is more appropriate than the Poisson model.

→ **Deviance criteria** is typically used in generalized linear model comparisons. Tables 3.4 and 3.5 display both the "null" and "residual" deviances. To remind briefly some definitions, "null" deviance states for the likelihood difference between the "null" model, taking only the intercept into account, and the "saturated" model assuming that each data point has its own parameters. "Residual" deviance (thereafter called Δ) corresponds to the likelihood difference between the estimated model and the "saturated" one. If the model is correctly specified (i.e. that the GLM assumptions hold), Δ should follow a Chi-Square distribution with $n-p$ degrees of freedom, n being the number of observations and p the number of estimated parameters (intercept and dispersion parameters excluded). More accurately, if the model is correct Δ should not be far from $n-p$ (2 535 in our case) [17]. Whereas Poisson model residual deviance is very far from 2 535 ($\Delta_{Poisson} = 10326$), Negative Binomial model is quite close to it ($\Delta_{NegBin} = 2542.4$), showing that this model could be more or less considered as a "good" model to fit the data.

[19] Using notations from equation (3.2), $g(\mathbb{E}(Y|X)) = \log(\mathbb{E}(Y|X)) = \beta' X$ in these cases.

[20] The correlation between claims count and $\log(\text{Traffic} \times \text{Portfolio size})$ amounts to nearly 60%.

[21] Codes : '***' states for a 0.1 % significance, '**' for a 1% significance, '*' for a 5% significance and '.' for a 10% significance

[22] Likelihood ratio test-statistic formula reminder : $D = 2(\log(Likelihood_{NegativeBinomial}) - \log(Likelihood_{Poisson}))$

[23] Compare the 1 degree of freedom chi-square critical value at 0.5% significance level (7.879) to D.

All tests performed strongly favour the Negative Binomial model, corroborating that daily claims count is *conditionally* over-dispersed. Moreover, deviance study shows that Negative Binomial GLM is appropriate to model our data.

Covariate	**Influence over daily claims count per driver** *(all other things remaining equal)*	**Significance**
Year	+ 1.15%	***
Summer	+ 1.94%	**
Autumn	+ 8.42%	***
Winter	+ 10.10%	***
Monday	+ 15.67%	***
Tuesday	- 5.70%	***
Thursday	- 3.72%	***
Friday	+ 1.84%	**
Saturday	- 4.51%	***
Sunday	- 25.14%	***
Public holiday	- 18.49%	***
Autoregressive parameter	+ 0.13%	***
Air temperature	+ 0.12%	*
Wind speed	+ 0.92%	**
Visibility distance	- 8.56%	***
Cloudy	+ 2.38%	
Rain	- 6.72%	***
Fog	- 7.33%	***
Thunderstorm	+ 3.74%	
Snow	+ 26.32%	***
Null deviance	17 723 on 2 555 degrees of freedom	
Residual deviance	10 326 on 2 535 degrees of freedom	
AIC	27 004	

Table 3.4: Poisson Generalized Linear Model results

Covariate	**Influence over daily claims count per driver** *(all other things remaining equal)*	**Significance**
Year	+ 1.16%	***
Summer	+ 1.81%	
Autumn	+ 8.05%	***
Winter	+ 10.06%	***
Monday	+ 16.69%	***
Tuesday	- 5.52%	***
Thursday	- 3.60%	**
Friday	+ 2.01%	
Saturday	- 4.58%	***
Sunday	- 24.96%	***
Public holiday	- 18.70%	***
Autoregressive parameter	+ 0.14%	***
Air temperature	+ 0.10%	
Wind speed	+ 0.87%	
Visibility distance	- 7.66%	***
Cloudy	+ 3.14%	
Rain	- 6.29%	***
Fog	- 6.88%	***
Thunderstorm	+ 3.98%	
Snow	+ 27.21%	***
Dispersion parameter σ	38.40 (standard error : 1.44)	
Null deviance	4 523.2 on 2 555 degrees of freedom	
Residual deviance	2 542.4 on 2 535 degrees of freedom	
AIC	22 652	

Table 3.5: Negative Binomial Generalized Linear Model results

3.3.4 Limitations

Generalized linear models are undoubtedly simple to interpret and estimate, but they also entail some limitations.

First, GLM are not well-shaped for dealing with interaction parameters. Indeed, interactions should be inferred before model fitting and a lot of attempts must be carried out before finding the accurate ones. We could have performed a step-wise like procedure on GLM estimation to select potential interactions, but it is time consuming, while tree models for instance are very efficient to automatically point out such interactions.

Second, even if GLM are less restrictive than pure linear regression models thanks to the link function introduced (see equation (3.2)), they remain strongly connected to linearity as $g(\mathbb{E}(Y|X)) = \beta' X$ is linear additive. Data can also be modelled using *non-linear* [24] mindsets.

Finally, GLM approach is *parametric* [25], based on the assumptions that $Y|X$ has a distribution belonging to the exponential family, and that the link function g is *known* and bijective. But these assumptions can be called in question in many cases. Even if in this case, the Negative Binomial model seems to be appropriate, it may be of great interest to test some *non-parametric* modelling to challenge our results.

Among *non-parameteric* and *non-linear* models that may ease GLM limitations, tree models and especially Classification and Regression Trees (CART) models, as well as Random Forests or Gradient Boosting Machines (GBM, also named Generalized Boosted Regression Models) can be considered as fine alternatives.

3.4 Gradient boosting model

In order to check if modelling daily claims count using GLM framework is not too restrictive, we fulfil this study by fitting a Generalized Boosted Regression Model. After explaining briefly how GBM model works, we will study more precisely the estimation results and challenge them.

3.4.1 How does this model work ?

Regression trees are a great alternative to standard linear and generalized linear models. Indeed, a tree model is based on a recursive binary splitting algorithm that iteratively splits the population in two subsets within which the target variable distribution is almost homogeneous [26] till a stop-criterion. For further technical details about regression tree construction, please refer to [22]. Regression trees are generally very straightforward to interpret, as they are usually graphically presented and rely on if-then split rules over some input variables. In comparison, GLM models results have to be interpreted "all things remaining equals" that may lead to some understanding difficulties. Moreover, a regression tree automatically detects interactions, whereas they must be manually added in a GLM model.

Nevertheless, regression trees predictive power is usually weaker than GLM one. Indeed, they are generally less robust : the problem with a single regression tree is the high variance it suffers from. More precisely, it means that a same tree design may yield very different outcomes from one fitting to another. There are several ways to solve this problem, among which are the *random forest*, *bootstrap aggregating* (also

[24] Non-linear extensions of GLM can be found in Generalized Additive Models (GAM), in which the linear predictor is composed of a sum of unknown smooth functions of input variables which may be either parametric or non-parametric. Model estimation concentrates therefore on finding these smooth functions. The general form of this type of model writes : $g(\mathbb{E}(Y|X)) = \beta_0 + f_1(X_1) + f_2(x_2) + ... + f_m(x_m)$ where $f_j(.)$ are the smooth functions to be estimated.

[25] In fact, when dealing with numerous observations, GLM formulation is no more restricted to the exponential family. Some other distributions can be used such as Extreme Value distributions. But it means that the likelihood function has no longer a single form and must be estimated using non parametric techniques. A R package was developed to deal with these extensions, called "biglm".

[26] Subset *homogeneity* is assessed using a chosen loss-metric applied to each subset (Gini impurity, information gain or variance reduction are the most used criteria). The values computed are then combined to provide a quality measure of the split. At each iteration, the "best" split chosen by the algorithm is the one that lead to the lowest loss-metric.

called *bagging*) and *gradient boosting* algorithms. All these techniques are based on multiple tree combination known to strongly reduce variance. *Boosting* principle applied to regression trees can be defined by the following sentences [8] : "In boosting, the tree grows sequentially. Each new tree is fitted on the residuals from the previous tree, which is strongly different from bagging in which trees are simply averaged".

Boosting main principles

A classical regression model aims at predicting a target variable y from a set of p covariates $x = (x_1, ..., x_p)$. It boils down to find a function of the vector of covariates $f(x)$ so that $\mathbb{E}(y|x) = f(x)$. $f(x)$ is estimated by minimizing a loss function that writes $L(y, f(x))$.

In a GLM framework, $f(x) = \sum_{j=1}^{p} \beta_j x_j$, while in a generalized additive model the parameter β_j is replaced by a smooth function f_j such that $f(x) = \sum_{j=1}^{p} f_j(x_j)$. In a boosting framework, $f(x)$ is also additive but smooth functions grown *sequentially* and may depend on all the covariates and not on only one of them. To be precise, the estimation is done *sequentially* so that $f(x) = \sum_{t=1}^{T} f_t(x)$ [a].

Let $f_t(x)$ be the prediction function guess at iteration t : $f_t(x) = \gamma_t \times h(x, z_t)$, where h is a function depending on the covariates [b], a vector of parameters $z_t = (z_1, z_2, ...)$ and a weight γ_t. γ_t corresponds to the weight given to the t-th prediction guess in the overall prediction function f. Function h is more precisely a simple predictive model called "weak-learner", for instance a neural network or a regression tree in this case. For instance, when regression trees are chosen as the "weak-learner family" $h(x, z_t)$ writes : $h(x, z_t) = \sum_{j=1}^{J_t} c_{jt} \times \mathbb{I}_{\{x \in R_{jt}\}}$. [c]

Boosting method thus seek $(\gamma_t)_{t\in[1,T]}$ and $(z_t)_{t\in[1,T]}$ sequences that minimize M quantity such that [d]:

$$M = \sum_{i=1}^{n} L(y_i, f(x_i)) = \sum_{i=1}^{n} L(y_i, \sum_{t=1}^{T} f_t(x_i)) = \sum_{i=1}^{n} L(y_i, \sum_{t=1}^{T} \gamma_t h(x_i, z_t))$$

Solution to the former optimization problem is solved by the following algorithm called "forward stagewise additive modelling" :

1. Initialize $f_0(x) = 0$
2. For iteration t from 1 to T, repeat :
   - Get γ_t and z_t estimates out of the minimization of the updated loss-function

     $$\sum_{i=1}^{n} L(y_i, f_{t-1}(x_i) + \gamma_t h(x_i, z_t))$$

   - For each observation i, update the predictive function guess : $f_t(x_i) = f_{t-1}(x_i) + \gamma_t h(x_i, z_t)$
3. Finally, return the estimated predictive function : $\hat{f}(x) = f_T(x)$

More precisely, if we choose for instance a squared-error loss function [e], step 2 rewrites :

$$L(y_i, f_{t-1}(x_i) + \gamma_t h(x_i, z_t)) = (y_i - f_{t-1}(x_i) - \gamma_t h(x_i, z_t))^2 = (r_i - \gamma_t h(x_i, z_t))^2$$

where r_i corresponds to the *residual* for the i^{th} observation after iteration $t-1$. Therefore, for a squared-error loss-function, we fit at each iteration a weak learning model on the current residuals, which entails a fitting enhancement at each iteration.

[a] T states for the maximum number of iterations of boosting algorithm.

[b] Note that f_t depends no more on a unique covariate but can depend on all covariates.

[c] J_t states for the number of final regions R_{jt} in the t-th regression tree. And in this case, the splitting variables, split points, and terminal node predictions are parametrized by z_t.

[d] n corresponds to the number of observations.

[e] Other loss-functions can be selected to fit the model, depending on the type of the outcome variable : L^1-loss to look at quantile regression, Bernouilli likelihood when considering a binary variable ... The one used for count data and Poisson distribution is detailed section 3.4.2.

Minimization performed at the second step is generally quite difficult to carry out. A specific optimization algorithm is therefore used to obtain γ_t and z_t estimators. In *gradient boosting model*, minimization is based on a gradient descent algorithm released by J. FRIEDMAN in 1999 [11].This two-step gradient descent algorithm applied to regression tree weak-learners is detailed figure 3.15 [27].

In plain text, this algorithm seeks to improve the current model (i.e. the model derived from previous iterations) by fitting a regression tree on its residuals r_i [28]. This constitutes one of the main advantage of *boosting* against *random forest*. Indeed, *random forests* only average bootstrapped trees fitted at the same level, while *boosting* seeks enhancement at each iteration.

Tree-based gradient boosting algorithm

In the first step, z_t is estimated by fitting a weak-learner $h(x, z)$ to the negative gradient of the loss function. We thus move in the direction of the steepest descent. Least-squares is used to fit the learner.

In the second step, the optimal value for γ_t is determined given $h(x, z_t)$, i.e. how far we move in the direction of the gradient.

1. Initialize $f_0(x)$ to be a constant $f_0(x) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$
2. For iteration t from 1 to T, repeat :
    - Compute loss-function gradient additive inverse for i from 1 to a chosen K such as :

    $$r_i = -\left(\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right)_{f(x)=f_{t-1}(x)}$$

    - Fit a regression tree (the new weak-learner) to r_i by least-squares and get the terminal leaves R_{jt} for j from 1 to J_t [a].
    - For j from 1 to J_t, compute the best gradient descent step-size c_{jt} :

    $$c_{jt} = \arg\min_c \sum_{x_i \in R_{jt}} L(y_i, f_{t-1}(x_i) + c)$$

    - Update f_t value such as : $f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J_t} c_{jt} \mathbb{I}_{(x \in R_{jt})}$
3. Finally, return $\hat{f}(x) = f_T(x)$

[a] J_t states for the number of leafs in the t^{th} regression tree.

Figure 3.15: Tree-based gradient boosting algorithm detailed procedure

As GBM fitting method relies on iterative enhancement, results generally face strong **over-fitting** if model hyper-parameters are not well specified. More details about the notion of "over-fitting" can be found in the below frame.

[27] This algorithm corresponds to the one implemented in "gbm" R package used in this thesis.

[28] Residuals can also be *re-weighted* according to the fitting error in order to boost fitting on misclassified elements. This re-weighting procedure is for instance implemented in "AdaBoost" algorithm, but we did not use it in this study.

What is "over-fitting" ?

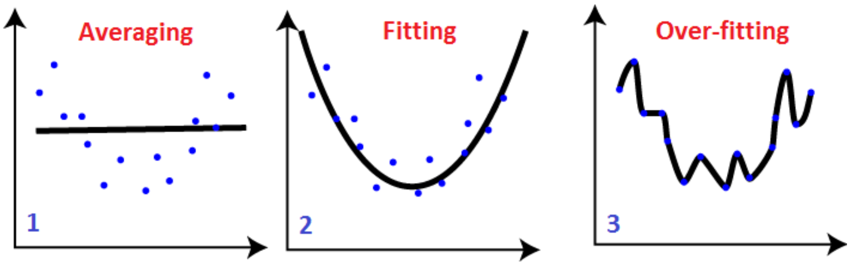


Figure 3.16: Toy example of over-fitting

In statistics and machine learning, we talk about "over-fitting" when the statistical model fitted ends up depicting random error or noise in place of the latent statistical structure of the data. It generally leads to counter-intuitive or even non interpretable fitting outcomes.

Figure 3.16 presents, on a toy example, the difference between a reasonable fitting and a situation of over-fitting. Blue points were randomly generated using a parabola equation with some noise. Chart number 1 shows the fitting curve (in black) of a linear model, obviously not adapted to the situation, as it only fits data average. On the contrary, chart number 3 fails to detect the underlying structure of data and fits only the noise introduced when generating the dataset. Finally, chart number 2 displays fitting results of a quadratic model that succeeds in detecting the latent data structure without capturing to much noise.

In order to cope with over-fitting curse in the GBM, several techniques were considered :

→ Choosing the optimal number of iterations, that is the **number of trees** generated, is essential to limit over-fitting. A "validation set" is extracted from the train set in order to compute the "validation deviance" of the model, which corresponds to the loss function value on the selected hold-out dataset. More precisely, it means that the model is fitted only on a chosen so-called "train-fraction" (thereafter called θ_{train}) of the train set, the remaining part being used to compute the "validation deviance". The "validation deviance" is expected to decrease gradually at each iteration (i.e. at each new tree grown), showing prediction improvement, till a minimum value before gradually growing again, showing the model enters an over-fitting stage. The optimal number of iteration matches the number of trees fitted associated to the minimum value of the "validation deviance" curve.

→ A **shrinkage** parameter λ can be specified to slow down fitting speed. Shrinkage reduces the impact of each additional fitted weak-learner $h(x, z_t)$ by capping its weight γ_t in the overall predictive function $f(x)$. Indeed, the intuition is that it is better to improve a model by taking many small steps than by taking fewer large steps. Thus, if one of the boosting iterations turns out to be erroneous, its negative impact can be easily corrected in subsequent steps. It can be shown that small values of λ ($\lambda \leqslant 0.1$ for instance) lead to significant improvements in over-fitting avoidance compared to the situation when $\lambda = 1$. Moreover, exploiting shrinkage in learning allows the decision-tree GBM to capture more continuity in the model effects, since multiplying the fitting steps lead to smoothing. Concretely, f_t update step in the tree-based gradient boosting algorithm is replaced by :

$$f_t(x) = f_{t-1}(x) + \lambda \times \sum_{j=1}^{J_t} c_{jt} \mathbb{I}_{(x \in R_{jt})}$$

→ **Sub-sampling** : Friedman proposed in 2002 to introduce some randomness in the fitting procedure [12] and proved that this additional step greatly improved the performance of the model. At each learning iteration, only a random part of the training dataset is used to fit the further weak-learner. The train dataset is generally sampled without replacement (method advocated in [12]). More precisely, it consists in randomly selecting a chosen fraction θ_{bag} of the train dataset just after the negative gradient computation step and in growing the further regression tree on this sampled dataset. Nevertheless, it should be used carefully, since if the number of points selected becomes to low, one might get a poor fit due to the lack of degrees of freedom.

Thus, at least six parameters values need to be selected to cope with model over-fitting : the number of trees fitted T, the maximum tree depth J_t, the minimum leaf size $Leaf_{size}$, the shrinkage parameter λ, the bagging and train fractions θ_{bag} and θ_{train}.

3.4.2 Results

In order to train our tree-based gradient boosting model, we split the data between a *train* set and a *test* set. As our raw dataset structure is time-based, we used a time split and kept observations before 2012 in the *train* dataset and put claims count recorded between 2012 and 2014 in the *test* dataset. This train/test data split will then enable us to compare GBM to previous Poisson and Negative Binomial models in terms of predictive goodness of fit (see section 3.5.1).

As the target variable is a count outcome, we chose to model $Y|X$ by a Poisson distribution [29]. The associated loss function writes :

$$-2\sum_{i=1}^{n}(y_i \times f(x_i) - \exp(f(x_i)))$$

Over-fitting avoidance was reached using the following set of parameters [30]. A seventh model tuning method imposing monotony on some variables was added to secure model consistency [31].

- T : 200 trees were fitted (see the appended validation deviance plot B.2.2).
- J_t : Each tree depth cannot exceed 10 interaction steps. It guarantees that learners remain "weak", i.e. not too intricate.
- $Leaf_{size}$: Each final leaf of the tree should have at least 10 observations in it. This ensures tree consistency, as it will less focus on outliers.
- λ : Shrinkage parameter was set to 0.01.
- θ_{bag} : 50% of the train dataset was randomly sampled at each iteration to grow new weak-learner on it.
- θ_{train} : 70% of the train set was used at each iteration to fit the weak-learner.
- Monotony was also imposed on some quantitative variables : it boils down to impose a monotony on partial dependence influences (see below). While preventing over-fitting, it enables easier results interpretation.

Table 3.6 gives, for each most influential covariate input [32], its "**relative influence**", that is more or less the percentage of model variance explained by it. To define explicitly the influence of a variable in the whole gradient boosting model, we must first define the influence of an input variable x in a single tree $\tilde{T}$. Considering that the tree $\tilde{T}$ has J splits, we look at all the non-terminal nodes, that is from the first tree split to the $(J-1)^{th}$. The influence of the input variable x in $\tilde{T}$ is given by :

$$Influence_x(\tilde{T}) = \sum_{j=1}^{J-1} I_j^2 \times \mathbb{I}_{(S_j=x)}$$

where S_j is the splitting variable selected at step j, and I_j^2 the empirical squared improvement assigned to the model as a result of this split. In plain text, this influence measure corresponds to the weighted number of times variable x was selected for splitting. To obtain the overall influence of the variable, influence is simply averaged over all trees of the GBM. The influences are further standardized so that they add up to 100%.

Covariate	Relative influence of the covariate
Visibility distance	25.80 %
Weather depiction	25.70%
Day of the week	23.97%
Air temperature	12.60%
Autoregressive parameter	4.43%
Season	3.35%

Table 3.6: Ranked most influential covariate on model variance

[29] Fitting a Negative Binomial distribution using a regression tree is much more complicated due to the over-dispersion parameter, thus not implemented in the R package we used in this study. But it would be interesting to seek an easy procedure to fit a Negative Binomial GBM.

[30] They were selected according to their performance out of a set of values sequentially tested.

[31] An example of GBM over-fitting on daily claims count dataset is appended figure B.3 page v.

[32] For clarity purposes, we display here only covariate that contributed to at least 3% of the overall model variance explanation.

We observe that the visibility distance, the weather depiction and day of the observation explain circa 75% of the variance explained by the model. These results are coherent with the generalized linear models fitted before.

Nevertheless, while pointing out the most influential variable, the previous table does not provide any information about *how* each covariate affects the response variable. Moreover, despite the simplicity of a simple regression tree, gradient boosting model combines hundreds of them, making results interpretation challenging. To better understand relations between the target variable and its covariates, *partial dependence plots* were created. These plots (see figures 3.17) show the influence of a selected covariate on the target variable after marginalizing out all other explanatory variables [33].

When using a GBM, one cannot really analyse the captured effects in a similar fashion to linear regression coefficients that are straightforward to understand. GBM results are more visual than quantitative but some orders of magnitude can be derived from these charts. The partial dependence plots show almost the same type of effects as in the Poisson and Negative Binomial models previously fitted [34] :

- When the visibility distance increases from 1 km to 2 km, daily claims count drops from nearly 125 to less than 95.
- Daily claims count on Sundays is far much lower than on other days : circa 80 on average against more than 105 on Mondays for example.
- Snowy and stormy days claims count records are much higher than for other weather.

Looking closer at the charts, the effects spotted are almost linear (see for instance the dotted slopes plotted in red on figure 3.18). This shows partly that GBM does not bring much more information on the data than what was derived from generalized linear models.

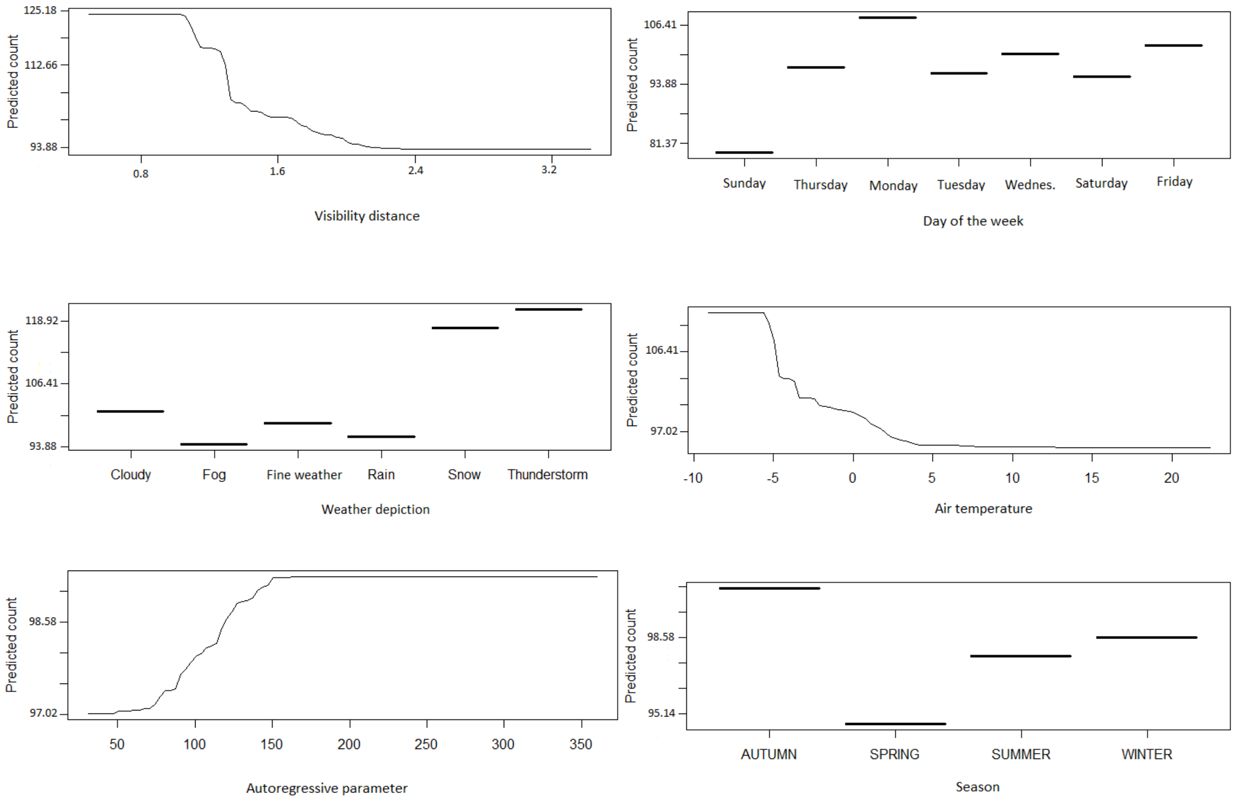


Figure 3.17: Partial dependence plots

However, one of our expectations when fitting this gradient boosting model was to catch some interactions between covariates that may have a significant influence on claims count (see 3.3.4). In order to check

[33] Detailed explanations about how such plots are computed can be found in the Appendices section B.2.2 page vi.

[34] Warning ! When the partial dependence curve turns out to be flat, it means that there were not enough observations in the slot to provide a meaningful partial effect.

that, we computed "interaction plots" : it takes the form of a heat map or grid chart crossing two input variables to highlight their joint influence on claims count. Unfortunately, the model fitted do not catch significant and disruptive interactions. For instance figure 3.18 shows the joint impact of visibility distance and weather depiction on the daily claims count. The evolution scheme of claims count according to the visibility distance is obviously the same whatever the weather reported. A slight interaction can be seen in the spread of evolution scheme which varies from a weather type to another, but it is quite far-fetched. These conclusions about interactions favour the hypothesis that GBM is not adapted to model our dataset.

There are in fact several possible reasons why a tree-based gradient boosting model may not be adapted in this case and fails to detect sensitive interactions. First the whole dataset is composed of circa 2500 observations only, and the train dataset on which the GBM was train is composed of nearly 1500 observations. It is very likely that our train dataset is too small to be modelled using such an iterative procedure. Indeed, bagging and validation steps may reduce greatly the number of observations on which trees are fitted, that can lead to misleading results. Second, weather and traffic data are quite heavily aggregated. Thus finding understandable interactions on the remaining signal after having removed all partial effects, is not likely to give many meaningful results. Finally, the covariate influences under scrutiny may simply be actually linear and the model fail to catch it because of its inherent complexity. A model comparison is carried out in the next section in order to better assess their performance.

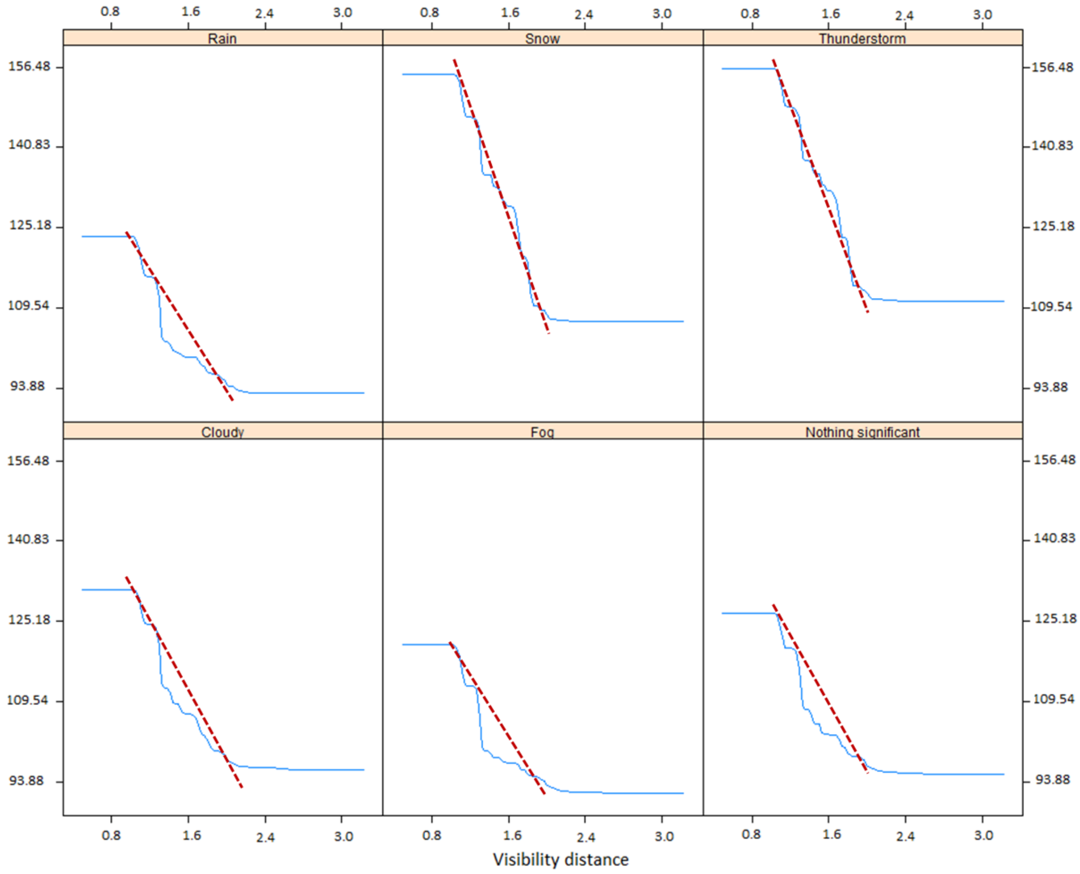


Figure 3.18: Interaction between weather depiction and visibility distance

3.5 Conclusion

Daily collision car crashes count was modelled using three models : a Poisson GLM, a Negative Binomial GLM and a Gradient Boosting Model. In all three models, time, weather and traffic data enrichment were proven useful to model daily claims count. We now want to challenge these models, first by comparing their prediction power and then by highlighting how they can be improved using telematics data resources.

3.5.1 Models comparison

As generalized linear models and gradient boosting models do not share the same type of performance measures, we chose to use Area Under the Curve (AUC) performance measure to compare them.

In order to compute it, we first retrained Poisson and Negative Binomial on a train set and computed predicted values on a test set [35]. Then the Lorenz curve of each model was drawn against the perfect model one and the random one (see figure 3.19). The Lorenz curve (also called "selection curve" in scoring techniques) represents the cumulative distribution function of the predicted probability distribution of daily claims count . More precisely, the curve is showing the proportion of overall claims count predicted by the model (Y-axis) assumed by the bottom $x\%$ of the observations (X-axis). The "perfect model" Lorenz curve corresponds to a situation in which all predictions made would have been exactly the actual value of the target variable. The closer to the perfect model curve the Lorenz curve under scrutiny is, the more performing the model is. Figure 3.19 shows that Poisson, Negative Binomial and gradient boosting model are quite close to the perfect model one, meaning that all three models are relatively performing. However, the Lorenz curves at stake are very close to each other : this does not really enable us to chose one model rather than the others. Moreover, the Lorenz curve associated to the gradient boosting model crosses Lorenz curves associated to generalized linear models. Thus looking at AUC alone can be misleading, as curves are not directly comparable.

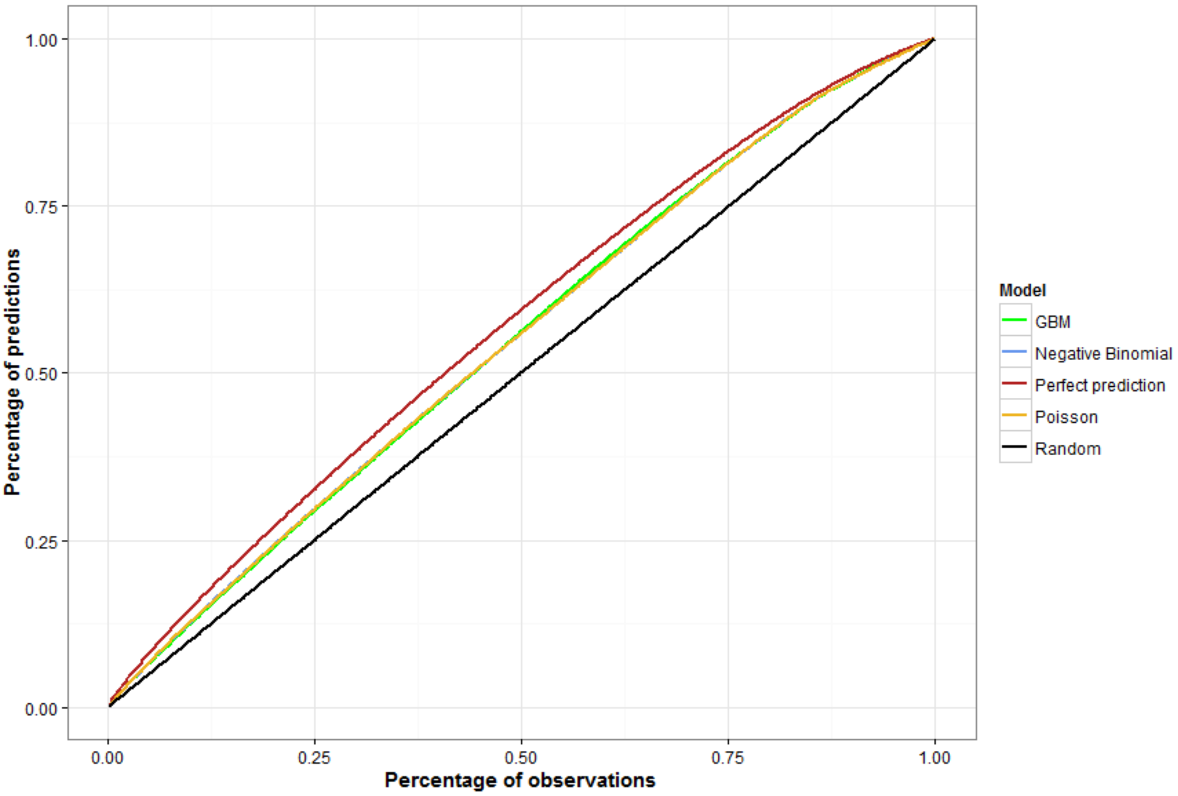


Figure 3.19: Selection curve of each model

To address this issue, the AUC performance measure was estimated. The AUC of a Lorenz curve is defined by the area between the random model curve and the curve at stake. Therefore, the AUC of each fitted model Lorenz curve was computed and divided by the perfect model AUC. Table 3.7 displays the results, which read : daily claims count is well predicted at $x\%$ by model k. We thus notice that, while having comparable predictive powers, the most performing model is the Negative Binomial GLM. It may be due to the fact that our data is actually conditionally over-dispersed, what the Negative Binomial model is the only one to catch and estimate [36]. GBM lowest score is also worse mentioning, as it means that parametric

[35] The same train and test sets as in the GBM fitting were used to enable meaningful comparison of models.

[36] The GBM model is indeed also based on a Poisson distribution hypothesis.

generalized linear model may be more adapted to model daily claims count than non-parametric and non-linear one.

Model	Percentage of perfect prediction AUC
Poisson model	67.67%
Negative Binomial model	68.08%
Gradient Boosting model	67.48%

Table 3.7: AUC comparison of models

3.5.2 Lessons learned from this study

This study enlightened how useful temporal features and external data can be to explain and predict daily collision claims count. Weather and traffic external data in particular may be considered by insurers as extremely valuable data to carry out their future motor risk analyses, all the more so as these datasets become more and more available thanks to open data movements. This type of temporal analysis of motor risk can be profitable for both the insurer and the customer from at least two point of views, if it is carried out on a regular basis :

- First, it can be used as a day-to-day risk management dashboard by the insurer. Anticipating daily collision claims count may help rationalize claims handling costs and reporting.

- Second, these studies can be used as an asset by the insurer to promote road safety to motor customers. Indeed, one can imagine a mobile application launched by an insurer that would warn drivers about the risks they face if driving that day, according to the traffic and weather situations encountered : for instance "since today is a public holiday, you may face less traffic, but collision hazard is high because of the snowy weather forecast". From the insurer point of view, it is a way to both reduce their claims charge and enhance road safety.

3.5.3 Why focusing on time-based data only is however restrictive ?

Nevertheless, analysing car crashes risk from a pure time-based point of view is restrictive. Indeed, as enlightened section 2.4, car crash risk depends a lot on geographical position and precise crash location surroundings. One of the drawbacks of this study is that it relies on country aggregated data, entailing averaged effects capture. In order to get more accurate insights about car crash risk, temporal *and* geographical aspects should thus be both considered when looking at car crash risks. This is why telematics data are tremendously valuable, as they record both precise time and location for each trip driven. Car crash risk can therefore be scrutinised at the local level (road level for example) and related to traffic and weather data specific conditions on a precise date. Moreover, telematics data will enable the insurer to better know to which actual traffic exposure his customers are facing on each specific road. This is one of the great advantage of telematics data compared to traffic data from watch organisations that rely on a network of static counters that cannot be enough widely spread to cover all types of roads.

In the next chapters, we will explain first how telematics data can be handled and exploited to make them valuable in the eyes of an insurer, and then use them to analyse car crash risk in Switzerland at a local road level from both a geographical and temporal point of view.

PART 4

Working with telematics data : "Big Data" at stake

In the previous chapter, we tried to explain daily claims count evolutions from a pure time-based point of view. We now want to add a spatial dimension to our car crash risk analysis. As mentioned section 2.4, car crash risk profile may change a lot from a region to another, and we assume that it may even be the case from a road to another. In order to test this hypothesis, we thus need to have access to road-level data. Moreover, this dataset must be at least dated to perform space-time analyses. Telematics data are therefore obviously relevant to carry out this type of statistical analysis. Furthermore, they are suited to model the insurer own risk as it reflects his own customers trips.

In this chapter, we will detail the procedure developed to make telematics data valuable from a statistical point of view and suitable to our goal. This section aims at explaining the difficulties an actuary may face while using telematics data to carry out his risk analyses. The next chapter will concentrate on model design.

4.1 Overview of the data processing designed

The procedure used to process and clean telematics data is composed of three stages that will be first introduced as a whole, and then detailed step by step. Telematics data were first completed using a "routing" technique. Then, the outcomes of the "routing" phase were exploited to define road segments and compute actual traffic flow on a road-level. Finally, traffic flow data was enriched using external data to create meaningful features that may explain car crash risk on a road-level.

4.1.1 Which type of datasets were exploited ?

The main goal of the data processing method we developed was to relate traffic data (exposure variable) with claims data (risk assessment variable) *on a road-level*. Traffic count was derived from the telematics database received from AXA-Winterthur, gathering all telematics recordings from January 2014 to March 2015. AXA Swiss telematics portfolio counts only a few thousands of customers. Claims count data was extracted from the FEDRO geo-located dataset we previously introduced (see section 2.4.1).

Telematics data format

Telematics device provided by AXA-Winterthur takes the form of a kind of "black box" directly plugged in the car. Whereas many other AXA entities adopted a smartphone digital solution for telematics, AXA-Winterthur favoured this type of recording device for its demonstrated better reliability. Data recording and initial cleaning is fully managed by an external service supplier.

Figure 4.1: Damaged road surface

Each time a telematics customer starts his vehicle, driving data is recorded on a regular timely basis by the "black box". It registers date and time, GPS locations the driver went through, which means the roads he drove on, accelerating parameters, cornering angles and severity speeding. As our aim is to analyse car crash risk linked to inherent road characteristics and time circumstances, we did not exploit accelerating parameters, cornering angles and severity speeding variables considered as behavioural driving parameters. Nevertheless, they may be integrated in further more detailed studies as they may include some statistical signal about road geometry and surface quality. Indeed, if a driver performs a lot of corners on a straight road, it may be due to a damaged road surface with a lot of potholes for example. Besides, if cornering angles are consecutively quite small, the driver may drive on a winding mountainous road. These are for the moment only hypotheses that must be confronted to driving expert judgements to be meaningfully integrated in a model.

As the volume of data automatically each millisecond recorded is too heavy to be stored by the device, only one GPS point every two kilometres was kept by the external service supplier, which entailed specific troubles while dealing with the data, that we detailed below. In addition to these every-two-kilometres points, so-called "event" points are also recorded each time the driver breaks heavily or performs a harsh cornering or manoeuvre. Nevertheless, as AXA-Winterthur telematics service is still in a pilot phase, "events" causes recording was biased at the time we received the dataset. We were thus not able to use these potentially useful pieces of information, and only exploited GPS locations associated to these "event" points to complete trips routes. We thus had access , for each trip, to time information and to GPS position at least every two kilometres. But even using this storage trick, the dataset at stake was nearly 40 millions observations wide. That is why we restricted our study to trips that went through four specific Swiss regions [1], which risk profiles were diverse and interesting (see section 2.4.4). Using these restrictions, the dataset dimension amounts to circa 15 millions observations, each observation being a GPS point recorded by the telematics device plugged in the vehicle.

Knowing the accurate locations through which each driver went, we can thus replenish the precise "route" the driver followed, that is the succession of roads he drove on. Gathering trips data from all AXA-Winterthur telematics customers, we can thus derive a proxy of the actual traffic flow bore by each road, and therefore a proxy of the actual exposure a driver faces when driving on this road.

Claims data format

Claims data was purchased from the FEDRO, and contains *only* latitude and longitude points for each bodily injury car crash recorded by the Swiss federal police between 2011 and 2013 (see details 2.4.1), which represents circa 160 000 bodily injury crashes. In order to make this dataset valuable, we had to relate GPS points to existing roads on a map, and match them with the traffic exposure variable derived from telematics data on a road-level. Then bodily injury claims were counted on each road, thus creating a risk assessment feature for each road. Indeed, from a pure geographical point of view, we assume that the more bodily injury crashes were registered on a road, the more inherently risky it is.

4.1.2 How were these data made statistically valuable ?

Figure 4.2 summarizes telematics data processing and enrichment designed to make it statistically valuable. It recapitulates each stage we had to go through to finally merge, for each road spotted, its exposure with its inherent crash risk. Two tasks were jointly carried out to get the so-called "ground database" in figure 4.2 page 59 :

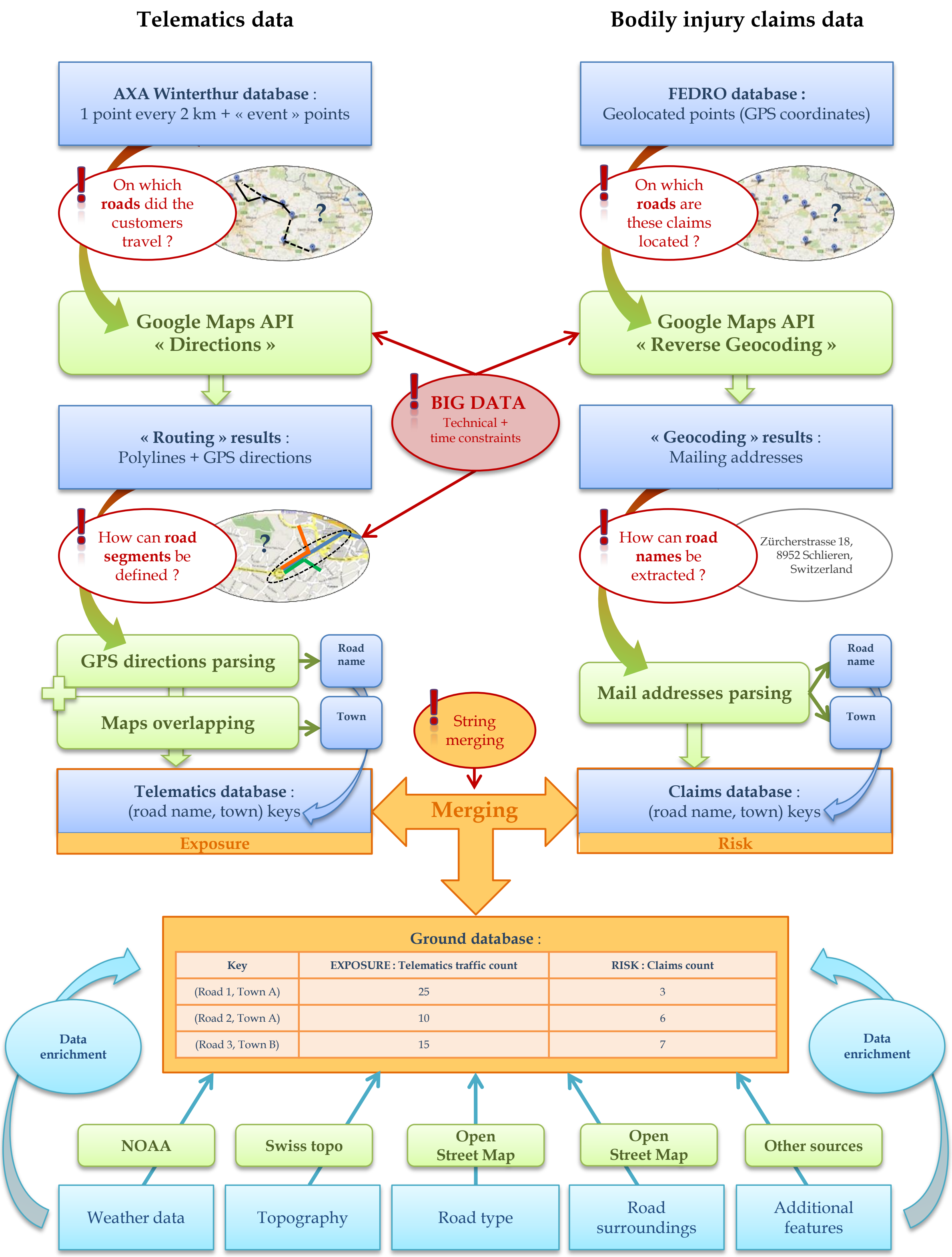
- **Telematics data processing** : First, GPS points may not be strictly located on an existing road, but can refer for instance to a building. This phenomenon is mainly due to interferences or to a lack of accuracy when sending and/or receiving position information from satellites. Therefore, we had to reposition available GPS points on existing roads (typically the closest road). Second, a point every two kilometres is not enough to derive the actual and complete route a driver followed (see figure 4.3), we thus had to use a "routing" machine to complete trips data (see section 4.2 for more details). Third, using these results, we defined more precisely what we call a "road", in order to be able to match claims data on it.

- **Claims data processing** : GPS positions had to be associated to an existing road in a step called "reverse geocoding" (see details below), in a similar fashion to what was done for telematics data. Then, the results had been cleaned to ease merging with the telematics database and thus relate exposure to risk variables.

Once this "ground" database had been constructed, a long lasting work of data enrichment was carried out to relate telematics data with external data, and especially with topographic and weather data, represented by the last light blue block in the data process pattern displayed figure 4.2.

[1] All trips counting at least one GPS point in one of the regions spotted section 2.4.

In the next sections, we will detail each ground stage of the telematics database construction specifying the technical tools we used and explain these technical choices.


Telematics data
Bodily injury claims data
AXA Winterthur database : 1 point every 2 km + « event » points
FEDRO database : Geolocated points (GPS coordinates)
On which roads did the customers travel ?
On which roads are these claims located ?
?
Google Maps API « Directions »
Google Maps API « Reverse Geocoding »
BIG DATA
Technical + time constraints
« Routing » results : Polylines + GPS directions
« Geocoding » results : Mailing addresses
How can road segments be defined ?
How can road names be extracted ?
Zürcherstrasse 18, 8952 Schlieren, Switzerland
GPS directions parsing
Maps overlapping
Mail addresses parsing
Road name
Town
String merging
Telematics database : (road name, town) keys
Exposure
Merging
Claims database : (road name, town) keys
Risk
Ground database :
Data enrichment
NOAA
Swiss topo
Open Street Map
Other sources
Weather data
Topography
Road type
Road surroundings
Additional features

Key	EXPOSURE : Telematics traffic count	RISK : Claims count
(Road 1, Town A)	25	3
(Road 2, Town A)	10	6
(Road 3, Town B)	15	7

Figure 4.2: Data processing pattern

4.2 "Routing" process

The word "routing" in our study must not be conflated with the usual meaning of "routing" systems in internet engineering stating for the process of moving data from a source to a destination. In the geographical framework of our study, it corresponds to a frequently used family of algorithms dedicated to map data completion [2]. More precisely, the aim of such algorithms is to link a series of isolated GPS locations taken as inputs by finding the most likely and shortest path that lead from one point to the following, using a known road-network, which usually takes the form of a map. Well-known route planning applications and software such as Google Maps or Mappy exploit this type of algorithms to provide their services.

4.2.1 Why using this technique ?

As mentioned in the data introduction, one GPS point every two kilometres is not sufficient to know exactly through which roads a driver went. Indeed, it is especially true when looking at urban areas : a driver may go from one end of a town to the other while travelling less than two kilometres and may have taken a lot of different routes since the street network is usually quite thick (think of Paris for instance). A "routing" algorithm is thus extremely useful to reconstruct complete trips. To understand, its usefulness on a toy example look at figure 4.3 : blue circles spot roads that would not have been taken into account without the routing machine, because no GPS point was directly located on them.

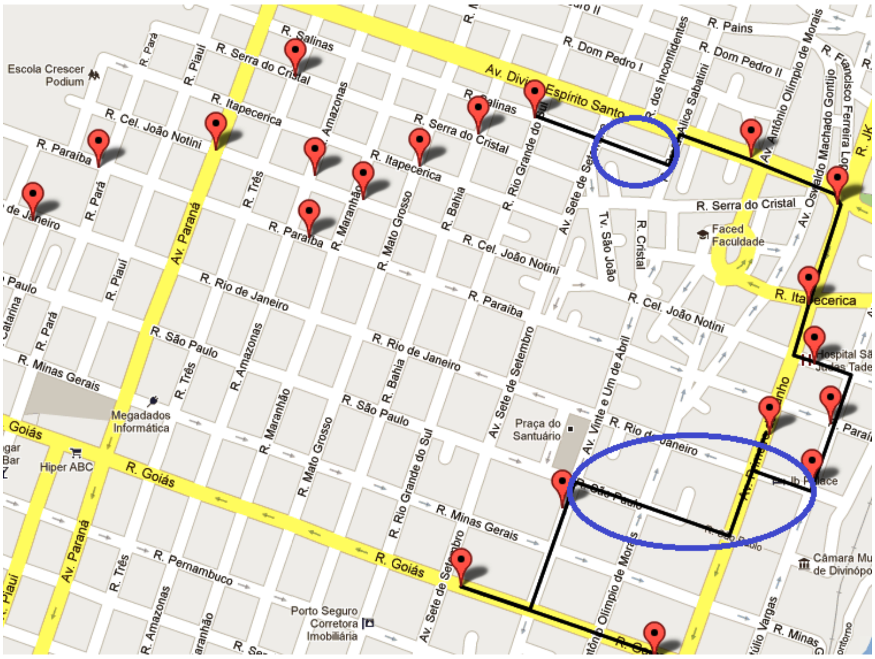


Figure 4.3: Toy example of a routing procedure

We must mention that keeping only one GPS point every two kilometres is attributable to the telematics service provider, which made this choice out of pure data storage criteria. Nevertheless, from an actuarial risk analyst point of view, it is not optimum. It indeed means great data loss, as expressed above, that he is compelled to balance by carrying out "routing" procedures for example. It would have been more easy and accurate to work on a database in which observations would have been kept every ten seconds for example.

[2] They are also often called "shortest path algorithms". One of the most famous one is based on network theory and is called Dijkstra's algorithm. More information about it can be found in the following document : http://math.mit.edu/~rothvoss/18.304.3PM/Presentations/1-Melissa.pdf

4.2.2 Practical use

Implementing by oneself a "routing" algorithm (or shortest path algorithm), is extremely difficult for at least two reasons. First, it relies on a road-network map that has to be sufficiently detailed to give meaningful results, map which is generally very expensive to purchase, and which becomes quickly out of date if not updated frequently. Second, assuming such a map is available, shortest path algorithms implementation is hard to optimize and accelerate : computation time may grow rapidly if the road network dimension is too wide. That is why we chose to use existing and performing "routing" solutions provided by external firms, and more precisely the Google Maps one as it was immediately available in AXA Global P&C.

Google Maps provides its services to firms through so-called APIs (Application Programming Interfaces), among which are the "Geocoding" API and the "Directions" one. It works on-line : the customer sends his request using a specific URL and gets back the desired results, generally coded using geographical data formats such as JSON or XML, on a web page the customer is free to save. More precisely, the "Directions" API is dedicated to route planning and thus uses a "routing" algorithm [3]. It retrieve the *most likely* [4] route a driver may have taken between several so-called "waypoints". Therefore, we created URL requests to send the circa 15 millions waypoint available in the telematics database at stake, and get back completed trips driven by telematics customers.

Using this Google Maps APIs implies nevertheless a lot of technical constraints, that are even more stronger as the data volume at stake is huge. These time and volume constraints have considerably impaired the data cleaning stage duration, as it took nearly three weeks to complete the "routing" process. These constraints must be taken into account while working with telematics data, as it may slow down potential real-time risk analyses processes. Among the main constraints entailed by Google Maps API usage, some are worth mentioning [5] :

- Each request to the API (that is the URL sent) can contain only a restricted number of waypoints, which means that if data points belonging to the same trip are too numerous, it must be split in several requests.
- The number of requests that can be sent within the same day is limited, as well as the number sent within a second, which means that the "routing" phase of telematics data has to be carefully planned to be done in a reasonable amount of time.
- Requests sending and results grabbing strongly depend on the internet access facilities available : when requests are automatically sent, proxy firewalls, internet disconnections and breakdowns, or low speed internet access may critically slow down the telematics data enrichment process.

Finally, "Directions" API requests outcomes were stored in JSON data format [6]. To derive useful and meaningful information out of it and integrate information in the database, a *parsing* phase had to be carried out. *Parse* data means analysing a string of symbols according to its specific syntax [7] in order to make it more understandable or to change its format.

4.3 Defining road segments

Google Maps "Directions" API retrieves, for each request, very specific results we introduce here. Using the waypoints sent through the URL request, it derives the most likely route linking all these GPS points. The outcome thus takes the form of a list of GPS device-like instructions, such as "Turn left on street A", "Go straightforward", etc., linked to more specific space-time elements. *Each* GPS device-like instruction (or "step") is completed by :

[3] Google somewhat quite uncanny about the algorithms its developers use, but the "Directions" API is said to be principally based on an improved version of the Dijkstra's algorithm.

[4] One may object that this type of algorithm only retrieve the *most likely* route linking two points and not the *actual* route taken by the driver. Yes, indeed, it is the case, but it brings much more information than considering isolated waypoints. Moreover, the most likely route computation takes into account road types and directions of traffic in order to be as accurate as possible.

[5] They may change according to the type of contract underwritten by the insurance firm.

[6] It corresponds to a relatively widely used data format that stores geographical data (latitude, longitude, mail addresses ...) proficiently and that is very easy to use when dealing with web pages.

[7] In fact, each data format or computer language has its own syntax (or "grammar") that proficiently shapes information displayed.

- A geographical item describing precisely the road segment on which the driver may have travelled on, called a "polyline". It corresponds to a series of tightened consecutive GPS points, which, if linked using straight lines, exactly match the road drawing on a map.
- An estimation of the driving duration and distance travelled. The computed duration takes traffic flow intensity [8] into account. The traffic flow intensity used corresponds either to the one registered at the time the request is sent, or to the one associated to given hours incorporated in the URL request. In order not to send too many privacy-sensitive data to Google API services, we chose not to include driving hours (though recorded by the telematics device) in our URL requests. Thus, durations retrieved from the routing procedure are associated to traffic flow intensity at the time we sent the request, which was done fully randomly in order not to bias the results [9].
- A depiction of the manoeuvre to execute to carry on the next route step.

To be more precise, a new GPS device-like instruction is issued at least any time the driver has to change direction or street. Additional instructions are issued when the driver needs to travel a long time straight-forward on the same road to remind him he does not have to change direction.

4.3.1 Problem statement

One of the main problem using these results is that a "polyline" item is extremely difficult to deal with, and that it heavily challenges the way a "road segment" is defined. Indeed, considering a specific street (for instance the "Avenue des Champs-Elysées" in Paris, see figure 4.4), a customer may drive on the first 50 meters of the street and then turns right, another may drive on the first 150 meters of the street before turning left, and another may arrive on the same street from the last major crossing and travel on the last meters of the same street. Even though the three latter drivers travelled on the *same* existing road (namely the "Avenue des Champs-Elysées"), the "polyline" outcomes will be different from one another, as they did not drive on the same portion of the street. Therefore, how can these results be reconciled to consider them as belonging to the same "road segment" ?

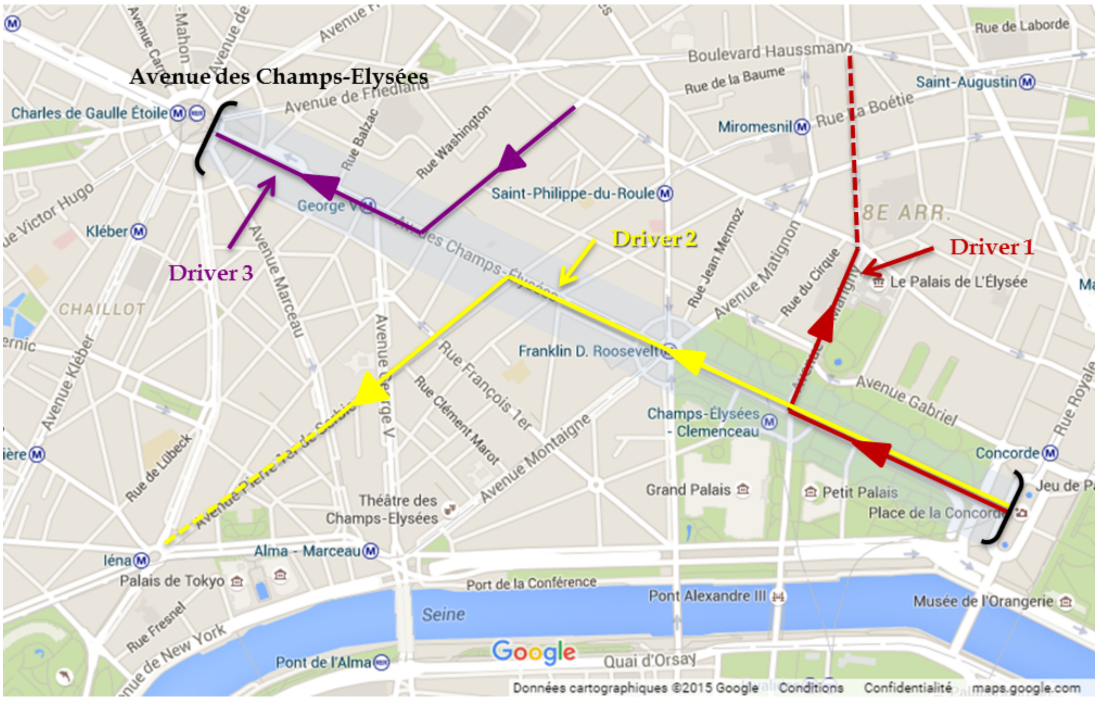


Figure 4.4: Situation in which a "polyline" solution is not efficient

[8] Real-time updated by Google services

[9] That is why driving durations on the same road can differ greatly, according to the time the request was sent : maximum and minimum durations were computed for each road. This is beneficial to our study as it enables congested roads spotting : if the spread between durations is heavy, it suggests that the associated road may be often congested.

4.3.2 Solutions

One of the solutions would have been to consider a "road segment" as a way located between two crossings, but this would have entailed at least two strong difficulties : first, it would have created very small segments (in towns especially) on which telematics traffic count would have certainly been too small to be significant ; second, it would also have created too long segments (think about highways for instance) that would have impaired surroundings variables consistency along the way. Another solution would have been to cross each polyline with all others contained in the database, and consider all polylines that have a portion in common as belonging to the same road. But first, this solution would have taken too much computing time according to the volume of data at stake, and furthermore, applying this solution to our toy example figure 4.4, it would have assigned a different "road segment" id to trip 3 even though all these trips go through the "Avenue des Champs-Elysées". The best solution, considering the issue, the data available and the time constraints we faced to complete this study, was thus to rely on road and street *names*. For example, all routes displayed in figure 4.4 go through "Avenue de Champs-Elysées", thus the telematics traffic proxy on this road segment amounts to 3 vehicles.

Nevertheless, deriving, for each route step, the name of the delimited road, was not an easy task. We first thought about "reverse-geocoding" (i.e. associate a written mail address to a given GPS point) each point of the polylines and derive the associated name from these results. But considering the amount of data at stake and all the constraints implied by Google Maps API usage, it would have taken far too much time to carry it out within the period prescribed. We finally chose to use the available written GPS device-like instructions to retrieve road segment names. Indeed, most of the instructions contained the name of the road on which the driver had to turn, and when missing, it entailed that the drivers had to stay on the same road as before, enabling completion. We thus used text-mining techniques to parse these instructions and get the names wanted.

At this stage, the database is thus composed of all the GPS device-like steps associated to each trip, filled in with road names associated. The final challenge laid in the fact that, while knowing the name of the road driven on, no variable indicated the town in which it was situated and though it is a critical piece of information especially needed to match locally external data. Indeed, some road or street names are widely used in the same country and may be found in hundreds of cities (think for instance of "Avenue de la République" in France). To solve this problem, we used an administrative map delimiting precisely each Switzerland city boundaries [10], and overlapped it with the GPS points contained in each "polyline" item. The city polygon that contained the greatest number of points belonging to the same polyline was considered to be the sought city polygon, and the city name associated was kept to complete the "road segment" id of the polyline.

Finally, the enriched telematics database is composed of all the GPS device-like steps associated to each trip, and completed with a "road segment" ID composed of a street/road name and the name of the town in which it is located [11]. To compute the traffic proxy by road segment we thus simply counted the number of distinct trip IDs that went through each road, using a "group-by" procedure. The "ground" database (see figure 4.2) we will now work on is thus composed of the road segment id variable, that identifies each observation, and a traffic proxy variable that reports the number of telematics trips that went through it, that is the "exposure" variable we looked for.

4.4 Merging with claims data

Bodily injury car crashes data contained *only* latitude and longitude points defining the location where the crash occurred. Examining the way we defined road segments, we had to retrieve road names and towns for each claim location latitude-longitude tuple, in order to be able to match them with telematics data.

The process of linking a given GPS point (latitude-longitude tuple) to its associated mail address (that is a series of words) is generally called "reverse geocoding" in the geographical field. It implies having access to a very accurate and wide map. That is why we also used convenient Google Maps services, namely the

[10] This type of administrative maps can be freely downloaded on GADM website : http://www.gadm.org/

[11] This key identifier is thus a *character* variable.

"Geocode" API [12]. As the claims database was not heavy (only nearly 160 000 observations), it did not took too much time to compute, but the constraints faced were the same as in the case of the "Directions" API (see section 4.2.2).

The resulting database was composed of the tuple latitude-longitude associated to its mail address. But in order to fully match it to telematics data, we had to clean these mail addresses and extract only useful pieces of information from it, that are road name and town. Indeed, a mail address contains a lot of useless items considering our goal : a postal code or a house number number for example. Once the cleaning phase was over, the claims database contained a similar "road segment" ID key to the one created in the telematics database. Finally, the "risk" variable was computed by counting the number of claims located on the same road segment.

Telematics and claims database thus both have road segment IDs as key identifiers. In order to get the "ground database" sought, they are merged using this key identifier. The main limitation of this merging process lies in the fact that it tries to match *character* variables together, whereas *character* variables are the most difficult variables to deal with. Indeed, the least misspelling, or changing accent may impair the matching process. And since the key deriving process was not the same for the telematics and the claims databases, it may have failed for a few road segments, but we tried to control it as much as possible. At the end of the process, almost 25% of the roads travelled on by the telematics customers registered a positive bodily injury crash count.

4.5 External data enrichment

Our goal is however to better understand car crash risk at the *road*-level, taking into account both spatial and timely surrounding circumstances. The previously built database, only composed of an exposure and a risk variable, is thus not sufficient to carry out our analyses. External data was used to enrich it and create meaningful features considering our goal (see last "orange" block in the data processing pattern figure 4.2).

4.5.1 Speed and distance travelled at a road-level

The "routing" phase (see section 4.3) was not only used to complete trip routes, but also to retrieve some useful features about the roads driven on. Among them are the duration and distance needed to complete each driving step, estimated directly by the Google API. Furthermore, knowing both distance and duration, travelling speed can be easily retrieved.

Distance travelled on a road segment

As mentioned section 4.3.1, several drivers actually driving on the same road may not necessarily travel the same distance on it. It depends greatly on the route they follow (see the example figure 4.4). Thus, when aggregating routing data by road segment IDs, the distance travelled on the *same* road is not unique. Indeed, according to the route they follow, drivers may travel on a tiny portion of the road before changing direction or simply drive all along the road. For each road segment, we get a range of distance values, that can be widely spread. That is why we created two different specific features from the distance variable contained in the routing database. :

- **Maximum distance** travelled on a road refers to the maximum figure within the range of distance values associated to that road. It can be seen as a good proxy of the actual road length.
- **Minimum distance** corresponds to the minimum figure within the same range of distances. To our understanding, it can be interpreted as the minimum distance a driver can travel on a road before facing a crossing. Think for example of a motorway compared to a street in a thick urbanized area : you must generally travel more than one kilometre on a motorway before being able to change direction (that is to follow the next step instruction in the routing database), whereas you may change direction after only few meters in an inner city street. It can thus be seen as a proxy of the number of crossings a driver may

[12] Even if it is called "Geocode" API, it also performs the reverse operation.

face by travelling on this road : the lower is the minimum distance travelled, the more numerous those crossings there may be along this road.

Speed

For each routing step, an estimated travel duration was also computed by Google API. This estimation takes into account the traffic flow intensity recorded by Google at the time the request was sent to the API (see section 4.3). As it was randomly done, for each road segment a range of duration values was extracted. It can also be widely spread, both because of the distance computation mentioned above and because of the changing traffic flow. Dividing the distance by its associated duration, we retrieved the travelling speed. We designed two different features from these figures :

- **Maximum travelling speed** associated to a road (maximum figure in the range of speeds computed for a same road). It corresponds to a good proxy of the maximum authorized speed. Indeed, when no traffic congestion is forecast by Google services, Directions API yields duration estimations based on the maximum authorized speed on each road.
- **Speed spread** between the maximum speed previously computed and its corresponding minimum speed on a same road. If the spread is high, it means that the travelling durations estimated by Google services on a same road can be very different from one date/time to another according to the traffic flow intensity associated. It can thus be interpreted as a measure of the propensity of the road to be congested. The higher is the speed spread, the greater is the likelihood that this road will be heavily congested at certain times of the day.

4.5.2 Topographic data

The Swiss Federal Office of Topography (also called "Swiss Topo" [13]) is responsible for geographical reference data. It provides the description and representation of Swiss geographic spatial data through national maps, elevation and landscape models, satellite images, etc. An elevation model of Switzerland, freely available, records height and slope at a 50 meters grid level. We used it to enrich the ground telematics database with topographic features that will specifically characterise mountainous areas.

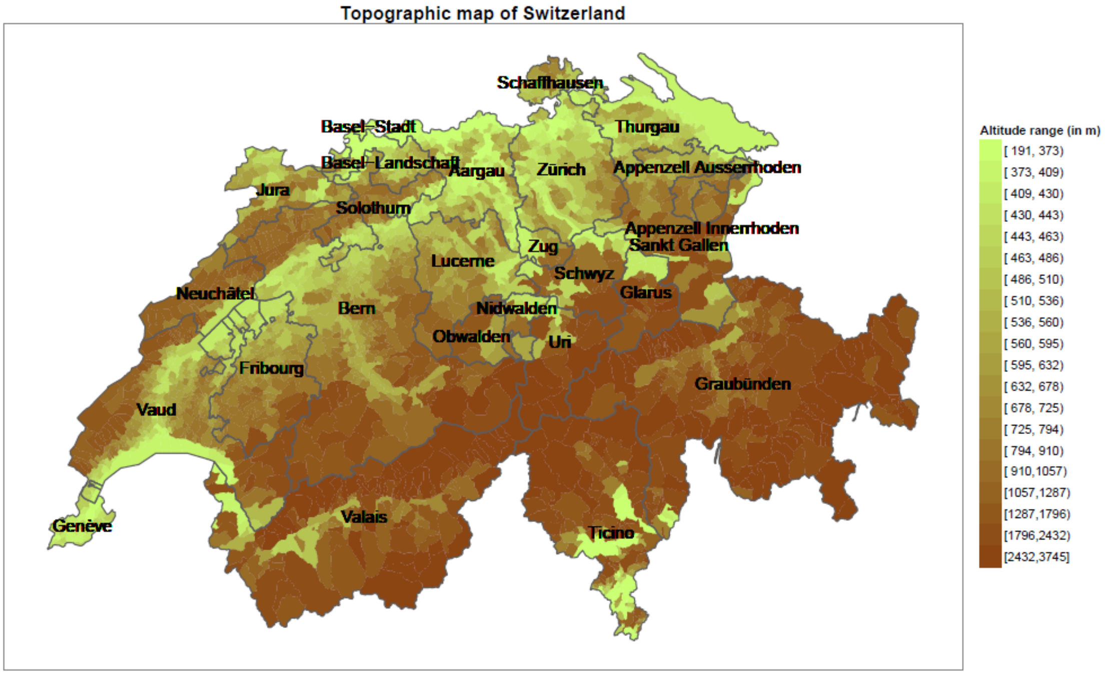


Figure 4.5: Height map of Switzerland

[13] More details are available at the following website : http://www.swisstopo.admin.ch/

The topographic database retrieved from the Swiss Topo elevation model is composed of a very wide range of geographical polygons [14] to which model-estimated **height** and **slope** are associated. Figure 4.5 shows the elevation model provided by Swiss Topo. Matching these data with the ground telematics database was carried out in two steps. First, topographic data was averaged at the city level through administrative map over-lapping : for each city, the maximum height and slope estimated by the elevation model were kept. Second, these features were added to the "ground" telematics database by city key merging, using the town name included in each road segment ID.

4.5.3 Weather data

Figure 4.6: Traffic jam on a snowy road

In order to enrich telematics data with weather features, we used the NOAA weather data introduced section 3.3.1. But this time, instead of averaging over all the country, we will use the weather stations locations. As the NOAA weather station network (see the map appended B.1) is not enough close-knit, the data features derived from them cannot differentiate directly two close roads, though our goal is to qualify weather features at the *road level*. That is why we decided to create weather features tightly linked to the actual telematics traffic computed after the routing stage. More precisely, we want to create *space-time* features that cross traffic flow at the road-level at specific date and time and the weather associated to the same time slot. For instance, features like "x % of the overall traffic flow that went on road A within the period under scrutiny was recorded on snowy days".

In order to link weather data with telematics traffic data that way, we proceeded in three stages :

1. As NOAA weather stations are relatively scattered, we first had to connect each road spotted by the routing process to its nearest neighbouring weather station [15]. The so-called "k-nearest-neighbour" algorithm [16] was used to do it, with $k = 1$ (i.e. we seek only one nearest neighbour).

2. We associated to each road segment the range of telematics trips that went through it, including the date and time at which the trips were recorded by the telematics device.

3. Knowing precisely on which day and at which time each telematics-equipped vehicle drove on the road segment, we merged the weather data [17] recorded at those date and time by the nearest neighbouring weather station of that road. We then derived, for each road, the *number* of telematics drivers who travelled on it by **foggy, snowy, rainy** [18] **and freezing weather**.

It would have been very useful to create the same type of crossed features for claims too in order to build one risk model per weather, to derive the inherent risk of each road under a snowy, sunny, etc. weather. It was however not possible, since the FEDRO crashes database is not dated. Nevertheless, even though it had been dated, the features created would not have been meaningful as the number of crashes at stake is too small : only 25% of the roads extracted count at least one bodily injury crash. Indeed, crossing crashes count according to the weather and time slot circumstances would have scatter statistical information and would have been detrimental to the results interpretation.

[14] Not automatically corresponding to a known administrative region.

[15] In order to get a unique GPS location for each road spotted, we geocoded the road keys created with the "Geocode" API.

[16] We could have recoded this simple algorithm, but its R implementation in the "FNN" package is very quick to compute. We thus chose to use it directly. See the "FNN" R package documentation for more details.

[17] We kept the temperature, weather depiction and precipitation amounts recorded by the stations. Precipitation amounts and temperatures were used to confirm or complete weather depictions recorded.

[18] We tried to split it between low,medium and heavy precipitations, but the statistical information ended up being too scattered, as the traffic count was automatically divided. We thus did not kept this split.

4.5.4 Surroundings depiction features

Enrich telematics data at the road-level cannot be done without qualifying the surroundings of each road segment. By the word "surroundings", we mean for instance the road type, the area classification between town and countryside, etc.

Road type

One of the surroundings features was retrieved from Open Street Map data [19]. Open Street Map is collaborative on-line open data geographical platform that gathers all the geographical information, quantitative and qualitative variables the volunteering web users want to upload for each country. Since it depends on web users contributions, a lot of useful variables may be available for some restricted geographical areas, while others would be "blank" of information or badly filled in. Information available in Switzerland can be considered as quite poor compared to France for example, which may be due to the lack of "open-data" culture in this country. Data is accessible via a specific API called "Nominatim", that works in the same way as the Google Maps APIs do : a specific URL containing the GPS location at stake is sent to the server that yields back a web page full of data. Among the variables available in Open Street Map Switzerland, we kept the road type one, as it was almost always filled. It distinguishes "primary roads" (major highways linking towns), "secondary roads" (roads that are not part of a major route, but nevertheless forming a link in the national route network), "tertiary roads" (used for roads connecting smaller settlements, and within large settlements for roads connecting local centres) and "other roads" (smaller roads, badly filled road types or unknown ones).

Region

Two other surroundings features were created from the routing results database. First, a region variable was created using an over-lapping of administrative maps. Each road segment was associated to the region in which it is located. As one of our goals is to assess the risk specificities associated to the regions selected (Zürich, Ticino, Geneva and Graubünden), we kept only these modalities in the variables. Other region categories were set to "not applicable" in order not to scatter the statistical signal contained in it [20]. We chose to take the region into account because we assume that drivers behaviour differs from one region to the other (see section 2.4). Not having included it would have meant that car crash risk is almost the same in all Swiss regions, which is contrary to our intuitions.

Manoeuvre

Second, routing process outcomes (see 4.3) contains the type of manoeuvre the driver must perform to access the next road of the route. The values encountered were summarized to three categories, namely "junction", "roundabout" and "turn". To access a same road, several manoeuvres can be performed by the driver according to its starting point and direction. Thus, in order to summarize the data at a road-level, we chose to keep the most common manoeuvre performed by the drivers to access this road. Several other ways could have been considered to sum up this variable, for instance taking into account the most dangerous manoeuvre performed to access it.

4.5.5 Note on time-based temporal features

In chapter 3, we pointed out several relevant time-based features to explain car crash count. Using telematics data, we want to score the risk associated to each road, embodied by the bodily injury crashes count reported on it. We could thus have included time-based features such as the day of the week, the month, etc. in our telematics-based model. Nevertheless, the telematics data records only cover a very short pe-

[19] Refer to the following website : https://www.openstreetmap.org/

[20] In the set of roads selected to fit the models (see next chapter), circa 40% of the roads were located in regions different from the ones selected. This is not surprising since the trips selected were not strictly limited to them, but may only go through them and come from a neighbouring region and end in another one.

riod (only one year and three months), compared to the daily claims count database that contained claims reported over six years. The time-based features created would have contained too weak statistical signal, and moreover impacts estimated would have been specific to year 2014 cause not averaged over several years. Data recorded on longer periods are necessary to carry out such more complex studies.

4.5.6 Note on "events" features

"Event" points are recorded each time the driver breaks heavily or performs a harsh cornering or manoeuvre. However, at the time we received the telematics database, the "events" were biased : the GPS point associated is relevant but the recording triggers were distorted. Indeed, the technical parameters thresholds that triggered point recording were not relevant. Thus we were not able to retrieve the categories to which they belonged (braking, cornering, ...). These elements will nevertheless be corrected in the coming months, and then integrated in our models.

4.6 Suggestions to improve telematics data collection and enrichment

Investments in accurate telematics and claims databases collection and storage, as well as in external datasets purchase, especially geographical and weather ones, will become more and more critical for insurance companies in the coming years. Telematics and more generally connected devices linked insurance data will indeed only bring value in actuarial risk analyses if they are properly enriched and feature engineered. For the moment, this enrichment represents a key sticking point in the growth of such insurance products. Indeed, telematics data enrichment must be considered by actuaries as the most challenging and demanding phase of motor risk analyses using this type of data. Due to its volume, initial storage processing and accuracy, data cleaning and enrichment processes may take a long time to compute, and thus make telematics pricing or risk analyses quite difficult for actuaries to carry out.

Based on the experience acquired processing telematics and external data, we suggest here some potential sources of improvement, that may greatly accelerate and ease future telematics motor risk analyses :

- **Claims** reporting would be far more useful if ***both*** **their precise GPS location and date of occurrence** were available. Actual space-time risk analyses might thus be performed to improve day-to-day claims management and regional hazard risk profiles.
- Telematics data storage process must be conceived to ease risk analyses. For instance, we saw in this study that keeping one point every two kilometres impaired a lot telematics data processing. Actuaries, **external data service providers and car manufacturers** should work together to improve telematics data recording devices.
- Thanks to **"open-data"** [21] movements, more and more external datasets are available on the web. However, when such a culture is not well developed in regions where insurers may have settled some entities, it would be beneficial for risk analysts to build efficient partnerships between insurance companies and private data providers.

[21] Open data is the idea that some data should be freely available to everyone wanting to use it as they wish, without restrictions from copyright or other mechanisms of control.

PART 5

Roads risk scoring using telematics data

The enriched telematics database now relates each geo-located road segment (observations) to a range of specific parameters (variables) among which are exposure and risk parameters retrieved from external data. Our goal is to explain, and possibly predict, the hazardousness of road segments according to their *local* environment and traffic parameters, thus catching the "inherent" risk of road segments. After introducing the design of the chosen models, we will interpret the results obtained. Last sections are dedicated to project future developments and improvements.

5.1 Models design

Contrary to the previous risk analysis which tried to predict *daily collision* claims count, the goal here is to predict hazardousness materialized by *bodily injury* crash count on each road. Bodily injury crashes count is therefore the target variable Y of our further models. The perspective is slightly different from previous studies for at least three reasons :

- First, the bodily injury crash database used is not dated, meaning that we focus here on a *space* analysis of risk. Out goal is to qualify the "inherent" risk bore by each road segment. *Time* is taken into account through the weather features (see section 4.5.3).
- Second, *bodily injury* crashes do not automatically depend on the same risk factors as *collision* crashes (see univariate statistics section 2.3.3).
- Third, *bodily injury* crashes imply generally a heavier cost for the insurer than *collision* ones that generally end up in property damage only. Focusing risk study on them can thus be useful to reduce claims cost.

Covariates introduction

In order to demonstrate that space-time circumstances are relevant while studying car crash risk, we chose to work gradually, enriching our model step by step. More precisely, new types of features were added at each step to show their usefulness in explaining car crash risk at road-level : exposure and speed parameters first, then topographic ones, weather features and finally surroundings covariates. At each step, a predictive power measure of the model was computed to objectively assess each model performance. Table 5.2 introduces the range of features X included in each model tested. They were created according to the data processing depicted in the previous chapter.

Traffic count and maximum distance driven variables are assumed to be "exposure" variables. We indeed presume that the longer you drive on a road or/and the heavier traffic flow there is on this road, the riskier it is. Nevertheless, we also presume that these effects may change greatly from one location to another. Therefore, we chose not to include them as *offsets* in the model, but to keep them as simple covariates in the model, enabling thus estimation of their direct impact on the target variable. Furthermore, in order to better visualize, the impact of distance and traffic variables on the probability to record a positive claims count in the partial dependence plots, we applied a *logarithmic* transformation on them in the model. Indeed, these variables are widely spread with a sharp right tail : using the logarithm instead of the actual value of the variable shrinks right tail effects and ease results interpretations.

Traffic count distribution			
1-th quartile	Median	Mean	3-th quartile
11	27	158.4	107

Table 5.1: Distribution of traffic count in the database reduced to significant roads (circa 24 200 road segments)

Covariate	Type	Meaning	Unit	Baseline modality
Model 1 : Traffic, speed and distance features				
Log(Traffic count)	Quant.	Number of telematics trips registered that went through the road under scrutiny	-	-
Log(Minimum distance)	Quant.	Minimum distance driven on the road under scrutiny according to Google Maps routing	Meters	-
Log(Maximum distance)	Quant.	Maximum distance driven on the road under scrutiny according to Google Maps routing	Meters	-
Maximum speed	Quant. (Levels)	Maximum speed recommended by Google Maps routing on the road under scrutiny	km/h	-
Speed spread	Quant. (Levels)	Speed spread on the road	km/h	-
Model 2 : Model 1 + topographic features				
Maximum height	Quant. (Levels)	Maximum height registered in the topographic map in the town in which the road under scrutiny is located	Meters	-
Maximum slope	Quant. (Levels)	Maximum slope registered in the topographic map in the town in which the road under scrutiny is located	Percent	-
Model 3 : Model 2 + weather features				
Traffic snow	Quant.	Number of telematics trips registered that went through the road under scrutiny by snowy weather	-	-
Traffic fog	Quant.	Number of telematics trips registered that went through the road under scrutiny by foggy weather	-	-
Traffic rain	Quant.	Number of telematics trips registered that went through the road under scrutiny by rainy weather	-	-
Traffic freeze	Quant.	Number of telematics trips registered that went through the road under scrutiny when the temperature was below 0 Celsius degrees	-	-
Model 4 : Model 3 + surroundings features				
Road type	Qual.	Type of road identified on Open Street Map	-	Other
Manoeuvre	Qual.	Most recurrent manoeuvre advised by Google Maps to join the road	-	NA
Region	Qual.	Swiss Canton in which the road is located	-	NA

Table 5.2: Covariates of fitted road risk models

We decided to select only the reliable and meaningful observations of the enriched telematics database, that is the roads on which at least 5 telematics-equiped vehicles drove on between 2014 and 2015 [1]. Nearly 42% of the remaining roads, that amounts to circa 24 200 road segments, have a positive claims count (against nearly 30% for the complete database). We must acknowledge that because AXA-Winterthur telematics offer is still at its early stages, the number of telematics customers is relatively low (only a few thousands) and the telematics data recording period was very short, leading to relatively low traffic count values (see table 5.1). Thus, the outcomes of this risk analysis cannot be directly generalized. Nevertheless, this study lays a first stone in the field of telematics motor risk analyses, as the methodology may be reused on

[1] 5 was chosen in order to get a relatively smooth distribution of daily claims count.

telematics data recorded on longer periods and broader portfolio.

Graphical insights

Univariate charts were first plotted in order to get some insights about how the covariates chosen may affect bodily injury crashes count recorded on the roads selected. Figure 5.1 [2] shows for instance that primary roads seem to be more risky than other types of roads. Roads being located at a junction point (ramp or fork) may be more risky than other ones according to figure 5.2.

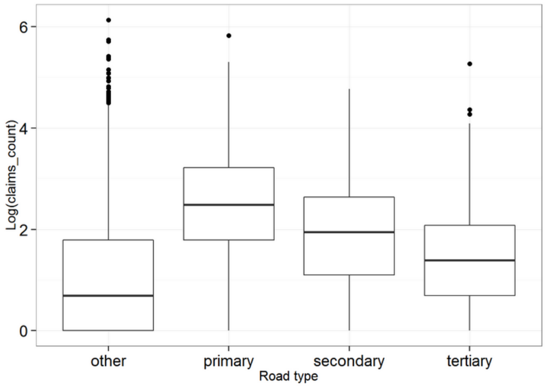


Figure 5.1: Boxplot of the logarithmic claims count according to the type of road

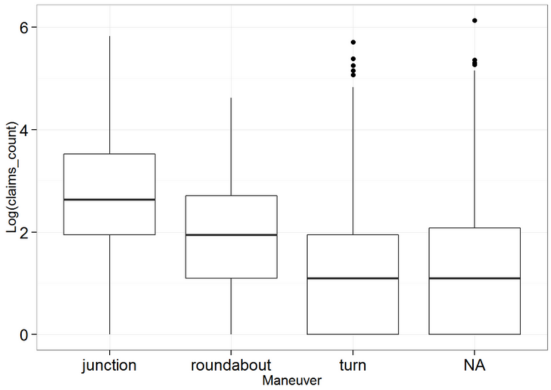


Figure 5.2: Boxplot of the logarithmic claims count according to the manoeuvre needed to join the road

Models selection and fitting method

In this chapter, car crash risk on a road-level is analysed from two different points of view : we first try to characterize the roads on which bodily injury crashes may occur, whatever the number of crashes reported on them (see section 5.2), and then attempt to understand which key factors may have an impact on the *count* of bodily injury crashes registered on each road (see section 5.3). In both case, we did not have any insight about how variables may interact with the risk feature we wanted to predict. That is why we decided to fit in both cases a Gradient Boosting Model, mostly praised for its adaptability (see technical details section 3.4.1). In the first case, we chose to fit a *Bernoulli* distribution as our target variable ends up being a dummy, equals to 1 if some bodily injury crashes were located on this road, 0 if not. And in the second case, a *Poisson* distribution was fitted, since the target variable is in that case, a *count* variable. Furthermore, in order to assess GBM predictive power in both cases, we also fitted Generalized Linear Models, a logistic and a Poisson one respectively.

As mentioned above, we chose to gradually add features in the model design. So, for each issue analysed (roads classification and bodily injury crashes count prediction), four "step models" were fitted (refer to table 5.2 for model details about each model specification). In order to specifically catch space-time covariates influence on car crash risk, the overall statistical design of the models does not change at each step : the range of covariates only is widen.

According to the statistical problem at stake, predictive power assessment of the models was done differently. When modelling the dummy variable, Gini indices were fitted from ROC curves, whereas when looking at bodily injury crashes count, the percentage of perfect prediction AUC was computed.

Finally, GBM fitting [3] were done on a train dataset composed of 60% of the significant roads randomly chosen. Models performance was computed using a test dataset of the remaining 40% of the roads.

[2] Logarithmic distribution of the bodily injury crashes count was chosen in place of the raw one in order not to overwhelm bottom meaningful influences due to outlier points.

[3] For both models, the over-fitting avoidance parameters chosen are : $T = 200$, $J_t = 10$, $Leaf_{size} = 10$, $\lambda = 0.01$, $\theta_{bag} = 50\%$, $\theta_{train} = 70\%$, and monotony was imposed on several variables.

5.2 Roads classification : On which type of roads do bodily injury crashes occur ?

The first gradient boosting model is designed to predict if a road is "risky" or not, that is associated to a positive claims count or not. We will first analyse the estimation results of the final model (i.e. fourth so-called "step" model), and then examine the differences in predictive power of each "step model".

Results

Table 5.3 summarizes the relative influences of the eight most influential covariates on the variance explained by the model. We notice that all surroundings variables (road type, region and manoeuvre) are among the most influential covariate, ’"road type" being the most influential one with 31.90% of relative influence. Then exposure variables (speed spread, maximum and minimum distances) bring a great part of the statistical signal. We nevertheless notice that traffic count is surprisingly very merely influential, as it explains only 1.2% of the dataset variance. Finally, we observe that topographic and weather features do not catch much statistical signal.

Covariate	Relative influence of the covariate
Road type	31.90%
Speed spread	18.48%
Log(Maximum distance)	15.95%
Region	8.37%
Manoeuvre	7.38%
Log(Minimum distance)	4.89%
Maximum height	3.94%
Traffic fog	2.38%

Table 5.3: GBM - Roads classification : Ranked most influential covariate on model variance

In order to better analyse the impact of each covariate on the target variable, we computed partial dependence plots associated to the most influential variables (see figure 5.3 page 73). Several features are worth mentioning [4]:

- **Road type** : Driving on a classified roads, that is on a primary, secondary and tertiary one, increases the likelihood to be involved in a bodily injury crash by 50%.
- **Speed spread** : The greater is the speed spread recorded on the road, the greater is the probability to be involved in a bodily injury car crash on this road. It means that, when driving on an often congested road (speed spread above 50 km/h), the likelihood to be in a car crash increases by nearly 28%.
- **Log(Maximum distance)** : The longer the driver travels on the road, the higher is his probability to be involved in a bodily injury crash on this road. It increases this probability from 32% to more than 44%.
- **Region** : As expected (see section 2.4), Geneva, Graubünden and Ticino region are riskier regions than Zürich. Indeed, for example, if the road under scrutiny is located in Geneva region, the likelihood to be involved in a car crash risk on this road is increased by 42.9% compared to a road located in Zürich region. Nevertheless, we notice that Graubünden is not riskier than Ticino or Geneva, contrary to what we found while analysing car crash risk form a geographical point of view. It can be due to the fact that our geographical analysis relied on an aggregated traffic data recorded on major roads, whereas here it relies on telematics data which is far more accurate. Furthermore, only roughly 3.5% of the overall significant roads are located in Graubünden region. This result must thus be carefully interpreted.
- **Manoeuvre** : When the driver needs to go through a roundabout or to turn in order to join the road under scrutiny, the likelihood to be involved in a bodily injury car crash on this road increases by circa 50% and 33% respectively.
- **Log(Minimum distance)** : The lower is the distance to reach the further crossing on the road, the more numerous crossings are likely to be located on this road (section 4.5), thus the riskier it is to drive on it.
- **Maximum height** : Roads located in a mountainous environment are slightly more risky than the others. Nevertheless, the harsh peak at the begin of the curve can be due to so-called "edge effect", making thus this effect not very reliable. Furthermore, topographic external data were aggregated at the city level,

[4] These interpretations must be related with the explanations delivered in the previous chapter : see especially section 4.5.

which may have impaired its significance on the road level (section 5.4.2 for more details).

- **Traffic fog**[5] : The effect of the percentage of the overall traffic driven on the road by foggy weather on the bodily injury crash probability of occurrence is quite interesting, even if this variable is not so influential. As soon as a part of the traffic recorded on the road (even if this part is quite small), the likelihood to be involved in a bodily injury car crash increases by circa 7.5%.

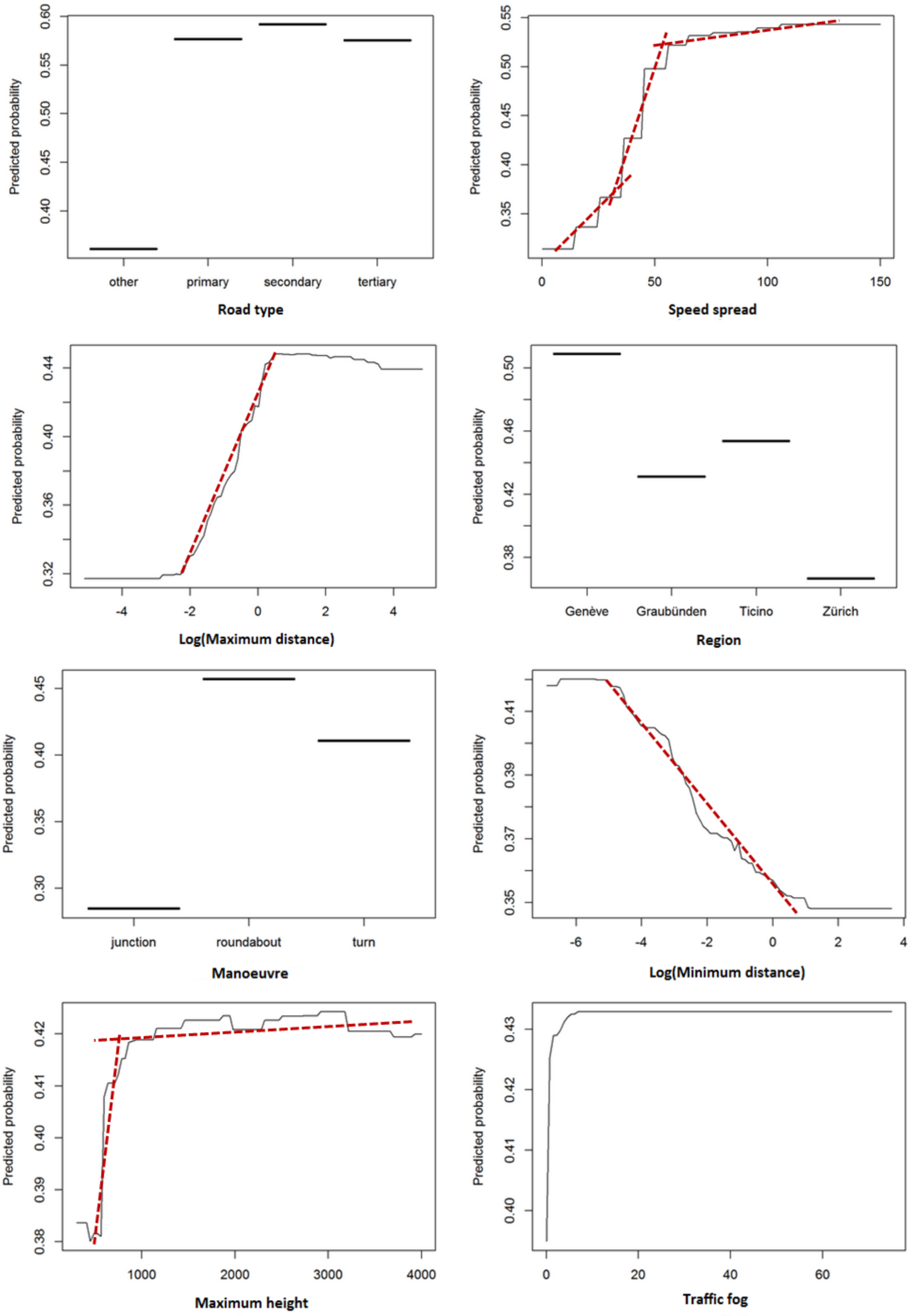


Figure 5.3: GBM - Roads classification : Partial dependence plots

[5] When the curve turns out to be flat, it is generally linked to a lack of data in the slot concerned.

The last three plots examined (log(minimum distance), maximum height and traffic fog) correspond to the least influential covariates in the model (explaining less than 5% of the variance). Thus, the interpretations derived from them must be taken with a grain of salt.

In the end, this model shows that weather and topographic features do not bring a lot of information when predicting if a road will be risky or not (i.e. counting at least one bodily injury claim on it or not), contrary to surroundings features. These conclusions are in line with our assumption that space circumstances must be taken into account to properly analyse car crash risk.

Models comparison

In order to check the relevance of topographic, weather and surroundings features when predicting the riskiness of a road, we computed the Gini index associated to each "step model". Results can be found in table 5.4.

At first sight, we see that the Gini index increases slightly at each step when adding new space-time features to the model. Because we computed this prediction quality measure on our test dataset, Gini index increase at each step cannot be put down to over-fitting, but directly to the predictive power of such space-time features.

Looking closer at them, we get a confirmation that surroundings features (included in model 4) seem to bring much more information than weather or topographic ones. Indeed, the Gini index raises by more than 8% between model 3 and model 4, against only circa 1% between model 1, 2 and 3.

In order to check if a Gradient Boosting Model was actually more adapted to fit the roads data derived from the telematics recording than a GLM one, we fitted a logistic GLM taking exactly the same covariates as the latter GBM depicted. The last line of table 5.4 is related to this latter logistic regression, which results can be found in the appendices figure C.1 page vii. Its Gini index amounts to nearly 48%, which is nearly 4 point lower than the GBM one, confirming that this type of model seems to be more adapted to model this type of phenomenon. However, the GLM performance may certainly improve if interactions and variable transformations spotted by the GBM are added manually as covariates. This modelling solution might be considered when industrializing such risk analyses.

Model	Gini index
Model 1	41.19%
Model 2	42.65%
Model 3	43.55%
Model 4	51.62%
GLM	47.79%

Table 5.4: Models comparison using Gini index

Limitations

Nevertheless, Gini indices values computed do not exceed 52%, whatever the model under scrutiny, which is not really high compared to standard levels. This relatively poor predictive performance may come from at least three elements.

First, the probability to be involved in a bodily injury crash may not only be due to space-time circumstances. Human factors, and especially the way of driving certainly have a strong impact on it. That is why we only capture a part of the statistical signal using space-time circumstances features.

Second, the claims database used was not dated, which prevented us from including actual time features in the model that could have be relevant.

Third, and this is certainly the most important bias, it may come from the database construction. Indeed, the exposure and risk variables were gathered using character keys merging (see figure 4.2 page 59), which may induce false "zero" observations. To be more precise, some road segments may be associated to a null bodily injury crash record, not because no claims were registered on it, but because the keys in each databases were strongly different while actually referring to the same road [6].

[6] When the difference between the keys is low, for instance qualified by a Levenshtein distance of 1, merging biases can be

Fourth, we must recall that both the telematics records and claims database used in this study were relatively small, which may have biased roads selection. Indeed, telematics exposure was recorded on one year and three months only, while the bodily injury crashes were reported on a three year period. These recording periods are not long enough to enable a robust road selection. For instance, some proven risky roads (i.e. roads with a high count of bodily injury crashes) may not have been driven on a lot by telematics drivers, which are mechanically excluded from the train database when selecting only significant roads. Using datasets collected on longer periods will certainly lessen these database construction biases and improve the prediction power of our models.

Because of these limitations, we decided to fit a *count* model on the data in the next section, in order to lessen the impact of database construction biases.

5.3 Bodily injury crashes count model : What makes a road riskier than another ?

We now focus on bodily injury claims *count* modelling. Roads will thus be more qualified according to their *risk severity* than according to their *riskiness* : the more bodily injury crashes occur on the road, the more risky it is. We try to predict the number of bodily injury crashes that may occur on each road using a count data statistical model, namely a Poisson gradient boosting model. This prediction will be then directly used as a **risk score** associated to each road.

Results

Table 5.5 displays the range of relative influence associated to the eight most influential features included in the model. The results are very different from the previous model fitted. We first remark that "exposure" variables (traffic count, and maximum distance) alone explain more than 40% of the variance, backing up our hypothesis that these variables are essential to scale the risk drivers take when selecting their route. Then, the type of road driven on still plays a great role in the prediction as it explains circa 11% of the overall variance. Finally, variables linked to the shape of the road (minimum distance to approximate the distance till the further crossing and manoeuvre) are useful to predict bodily injury claims count on the roads examined.

Covariate	Relative influence of the covariate
Log(Traffic count)	28.93%
Speed spread	11.71%
Log(Maximum distance)	11.57%
Road type	11.08%
Maximum height	9.56%
Log(Minimum distance)	6.39%
Traffic snow	4.74%
Manoeuvre	4.44%

Table 5.5: GBM - Bodily injury crashes count prediction : Ranked most influential covariate on model variance

Figure 5.4 page 77 gathers the partial dependence plots associated to the gradient boosting model targeting bodily injury crashes count :

- **Log(Traffic count)** : When the telematics traffic count exceeds roughly 400 cars ($\exp(6) \simeq 400$), doubling the traffic count on the road increases the number of bodily injury crashes located on it by nearly 150%.
- **Speed spread** : The partial influence spotted here is almost the same as the one spotted in the classification model : roads that are likely to be congested are riskier than others.
- **Log(Maximum distance)** : If the road is more than circa 1 kilometre long, the longer it is, the more numerous bodily injury crashes may occur on it. It corresponds to an intuitive geographical effect : the larger/longer is the area under scrutiny, the more numerous events may happen on this site.

corrected.

- **Road type** : Primary and secondary roads may record more than twice as many bodily injury crashes than smaller ones (tertiary and other categories).
- **Maximum height** : High mountainous roads seem to count circa one extra bodily injury crash compared to other roads.
- **Log(Minimum distance)** : When the distance till the further crossing increases, the number of bodily injury crashes predicted tends to slightly diminish.
- **Traffic snow** : As soon as a part of the traffic flow estimated on the road was driven by snowy weather, it raises slightly the number of bodily injury crashes recorded on the road (by nearly 5%).
- **Manoeuvre** : Roads that need to be joined through a roundabout or a fork seem to record more bodily injury crashes than the other ones.

The last two interpretations (traffic snow and manoeuvre) must be carefully taken into account as they are associated to poor influential covariates in the model (explaining less than 5% of the variance).

This model demonstrate that exposure variables (traffic count and road distance) must be taken into account when analysing car crash risk at the road-level. Moreover, context variables such as the height of the town in which the road is located or its type play a great role in roads risk assessment. We nevertheless see that, like in the roads classification model, weather features do not bring much information. It may however be due to the way weather features are constructed (see section 5.4).

Models comparison

As the models fitted in this section are not based on a dummy variable anymore, we had to rely on another prediction quality measure to assess their performance. We use here the ratio of AUC (Area Under the Curve) reached by the model under scrutiny over the perfect prediction AUC (see section 3.5.1 for more details). Table 5.6 summarizes the outcomes.

We first notice that the percentage computed increases while adding more space-time features in the Poisson GBM. Indeed, bodily injury crashes count is well predicted by model 4 at 68%, against circa 62% for model 1. Then, we see that the percentage of perfect prediction AUC remains the same between model 2 and model 3 : it means that the weather features created seem to be irrelevant to model bodily injury claims count at a road-level. It may directly come from the way those features were constructed, as in the previous model. Finally, we see that adding surrounding features increases the percentage of prediction AUC by circa 5 points.

Again, these outcomes favour the hypothesis that taking space-time circumstances into account to model car crash risk is relevant. Moreover, it back ups the idea that car crash risk can be modelled at a road-level, and that car crash risk is not only related to human factors pointed out by road safety enhancement campaigns, but is tightly linked to each road inherent characteristics.

A Poisson GLM was also fitted on the data to check if a simpler model than a Gradient Boosting one could be suitable. Results are available in the appendices page viii. Though almost all coefficients fitted are significant, the predictive power of this model is really low as the percentage of perfect prediction AUC reaches only 25%. It may be due to the fact that data is over-dispersed, meaning that the Poisson distribution is not adapted to the phenomenon we try to fit. If it is actually the case, a Negative Binomial distribution may have been more adapted (see discussions about the differences between those models in chapter 3 section 3.3.2). But looking closer at the partial dependence plots above, we notice that some of the effects spotted seem to be non-linear (speed spread and log(minimum distance) especially), favouring the fact that a non-parametric and non-linear model that automatically takes interactions into account, such as the Gradient Boosting Tree model is better adapted.

Model	Percentage of perfect prediction AUC
Model 1	61.96%
Model 2	63.27%
Model 3	63.27%
Model 4	68.08%
GLM	24.88%

Table 5.6: Models comparison using the percentage of perfect prediction AUC

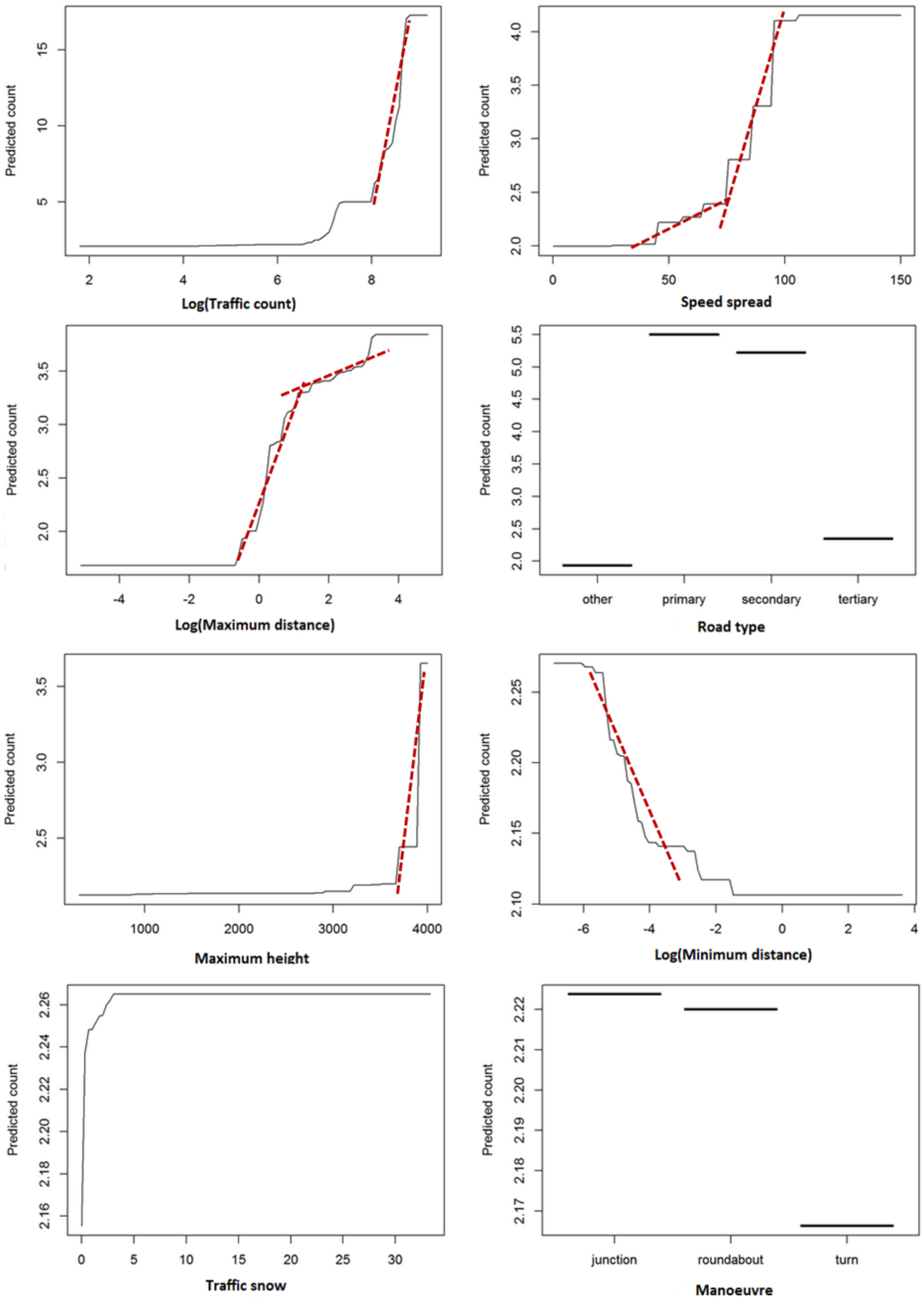


Figure 5.4: GBM - Bodily injury crashes count prediction : Partial dependence plots

5.4 Limitations

5.4.1 Model design

Gradient Boosting Models predictive power is generally quite high compared to more classical statistical models such as the Generalized Linear ones. It essentially hails from its fitting flexibility, especially when working with small trees as weak learners. Nevertheless, it has some drawbacks that need to be mentioned.

First, combining a lot of simple trees enables to easily catch hidden **interactions** between covariates. However, the way they are visualized on interaction plots can be challenged. Indeed, GBM algorithm does not take *constraints* into account when deriving interactions plots. For example, the covariates "minimum distance" and "maximum distance" are sorted : on a same road, the minimum distance cannot be larger than the maximum distance. The GBM algorithm still tries to infer statistical signal in value areas that are not mathematically possible. Figure 5.5 shows an example of this phenomenon. For instance, the right bottom area refers to tuple values that cannot be found in the database as they do not fulfil the mathematical condition needed.

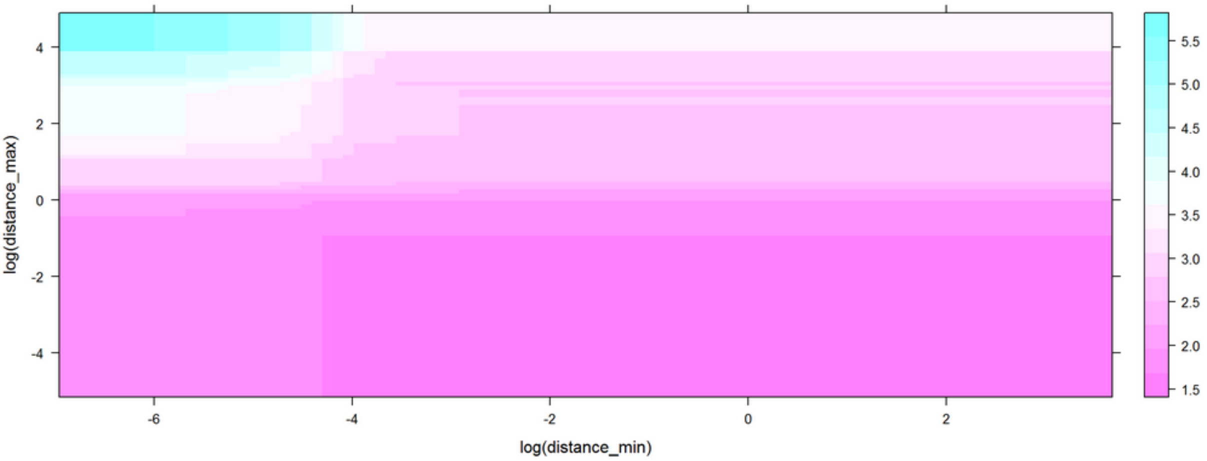


Figure 5.5: Interaction between distances

Second, this type of model, while being very efficient to carry out predictions, is **not easily readable**. Indeed, the interpretations are mostly retrieved by chart reading, which is not very accurate. On the contrary, Generalized Linear Models enable easy impact quantification but rely on strict statistical hypotheses that may often fail when fitting them on complex datasets. This drawback has to be taken into account while industrializing statistical and machine learning processes designed for risk analyses. The trade-off between an easy readability and a strong predictive power has to be balanced according to the purpose of the study. Before selecting any model, it is essential to clearly define those goals, by answering the following questions : what will be the final usage of the model fitted ? Do I prefer to predict well the target variable or to quantify some of the underlying phenomena ? What is worse : to under-estimate or over-estimate the phenomena ? To increase the rate of false positive or negative ? etc. It highly depends both on people to which they are dedicated and on people that will use it. For instance, if the previous models are directly dedicated to drivers in order to enhance road safety, it would be better to get a high predictive power. On the contrary, it would be better to create a more readable model if it is meant to be used daily by motor pricing actuaries. In this study, we chose to tackle the issue from a road safety point of view and thus value predictive power of the models fitted, more than readability.

5.4.2 External data weaknesses

The previous models interpretations showed that weather and topographic features may be poorly relevant to explain car crash risk at the road-level. However, their lack of significance certainly comes from the way those features were constructed, or from the lack of accuracy of the external databases used.

First, weather features had to be built by crossing NOAA **weather data** with telematics traffic count in order to get weather features specifically linked to each road, which is not ideal. It is mostly due to NOAA data processing : on the one hand weather data are not always filled in for all weather stations of the network that reduced the accuracy of the features constructed ; on the other hand, the NOAA weather station network is not tightly knit, meaning that a lot of roads are linked to the same weather station, that the data is **too aggregated**.
Ideally, we would have wanted to create weather features which are directly connected to the road thus enabling differentiation between a road and its neighbouring one. For example, if the weather data had been specific to each road examined, we would have created features such as a dummy indicating if this specific road is frozen or not, if there are puddles on it, etc. In order to get such *road specific* weather data, one can imagine some specific sensors directly plugged in the tyres of the telematics-equipped vehicle, that real-time record road temperature, presence of water on the road surface, etc., and which directly link these data to GPS positions reported by the telematics device.
Furthermore, it would have been more accurate for instance if they had been directly linked to crash data. If the bodily injury crash data had been dated, we could have created features such as the number of claims that were recorded on a snowy day, on a foggy day, etc., which may have brought more statistical signal in the models. Moreover, it would have been used to fit one roads risk model per type of weather.

Second, **topographic** features were derived from the Swiss topo model which, while being relatively accurate, do not associate specific features to each road. Height and slope measures are aggregated at the locality level, whereas two roads belonging to the same locality may have greatly changing slopes, especially in mountainous areas (think of Grenoble in France for example). We would have wanted to know the exact slope of each road, and its specific elevation.

This lack of accuracy of external open data can be compensated by purchasing datasets created by private firms. Nevertheless, these datasets remain generally very expensive, and is not always well adapted to the purposes of risk analyses.

5.4.3 Model on asserted risky roads

As mentioned above, even if we tried to lessen those biases and to chose the best solution at each step of the data processing, our dataset may suffer from construction biases. We acknowledge that the greater bias may be put down to the exposure and risk variables merging thanks to character keys. It indeed induced an ambiguous interpretation of a null claims count record : it is either due to a failure in the merging or to an actual absence of bodily injury crash reported on this road.

In order to get rid of this bias, we decided to create an inherent roads risk score on roads that count at least one bodily injury crash on them. It indeed implies that there was no problem merging risk and exposure databases (see data processing design figure 4.2 page 59). We thus fitted a Poisson GBM on the remaining roads to model their "degree of riskiness" [7]. We did not proceeded step by step this time, but included directly all the features introduced section 5.1 as covariates in the model.

Covariate	Relative influence of the covariate
Log(Traffic count)	27.68%
Log(Maximum distance)	22.33%
Maximum height	12.63%
Road type	8.36%
Manoeuvre	6.77%
Speed spread	5.79%
Traffic freeze	4.28%
Traffic rain	2.51%

Table 5.7: Bodily injury crashes count prediction on risky roads : Ranked most influential covariate on model variance

[7] Indeed, by restricting the dataset to roads that recorded at least one bodily injury crash on them, we restricted the study to already classified risky roads.

Table 5.7 presents the eight most influential variables of the model and the relative influences associated. We notice that exposure variables (traffic count and maximum distance) explain circa 50% of the variance. It means that, when looking at already risky-classified roads, the exposure parameters play a greater role than when mixing risky and non-risky roads. Maximum height is more influential in this model than in the previous one as it explains almost 13% of the overall variance, whereas the speed spread is far less influential than before. Finally, we remark that two weather features (traffic rain and traffic freeze) stand out as the seventh and eighth most influential variable, at the expense of the minimum distance. The fact that the minimum distance is not significant any more in this model (in which database construction biases were removed) potentially means that this variable caught in fact database construction biases in the previous model.

Figure 5.6 page 81 gathers the partial dependence plots associated. Almost the same interpretations as for the Poisson GBM fitted on all significant roads can be derived from these plots. Some slight differences are worth noticing :

- **Maximum height** : The impact of mountainous environment on bodily injury crash count is greater in this model than in the previous one. Indeed, roads located in high mountainous localities (beyond 3500 meters high) count twice as many bodily injury crashes as the other roads.
- **Road type** : A clear order between roads of different types is derived by the GBM : primary roads are more risky than secondary ones, that are also more risky than tertiary and other ones. Moreover, the difference in bodily injury crash risk is heavy as primary roads record twice as many crashes as tertiary and other types of roads.
- **Manoeuvre** : According to this model, roads that need to be joined through a ramp or a fork are more risky than the other ones (25% more bodily injury crashes are recorded on these roads). One may interpret it as the effect of driving blind spots. Indeed, drivers who need to join a major road through a ramp are very likely to be involved in severe accidents with heavy goods vehicles, which have a strong lateral blind spot on the passenger's side, or with other cars. They have to cope with a strong blind spot too, while having to drive fast to enter traffic safely on the major road.
- **Weather features** : There is admittedly an increase in the number of crashes as soon as some part of the traffic flow was recorded under rainy of freezing weather. Nevertheless, these effects are limited as the number of crashes counted on these roads increases by only 2 to 2.5%.

Finally, this model is performing as its percentage of perfect prediction AUC reaches 64.68%. It confirms that a great part of car crash risk associated to each road can be associated to its inherent exposure and surroundings characteristics.

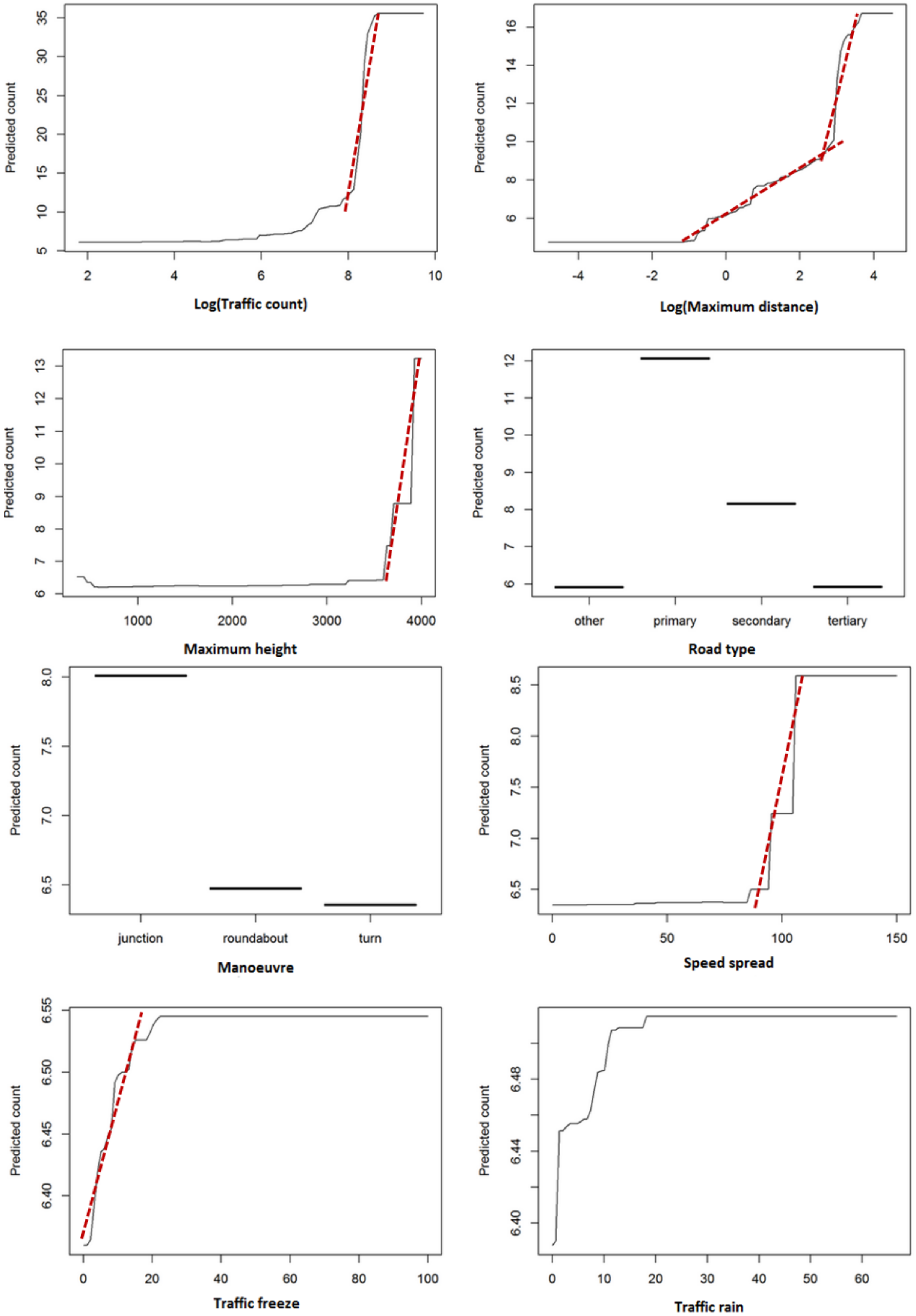


Figure 5.6: Bodily injury crashes count prediction GBM on risky road : Partial dependence plots

5.5 Further developments of the project

The risk analyses carried out on telematics data are still in the early stages. They first need to be improved technically, by testing other models, enhancing data quality and consistency at the road level and performing analyses on longer telematics records. These risk studies must also find their place between road safety enhancement programs and insurance business products.

5.5.1 Technical suggestions to improve the models

Looking at the potential richness of telematics data, several other statistical models and machine learning algorithms may be useful to derive meaningful insights about car crash risk to feed actuarial motor risk analyses. For instance, roads classification can be performed using statistical models different from the Gradient Boosting Machines : Support Vector Machines based on kernels, Neural Networks or even Naive Bayes classifiers must be considered. As far as road risk intensity is concerned, it can be examined differently according to the weather and/or traffic flow at stake. One can imagine a regime-switching time series model targeting the bodily injury crashes count recorded on each road, which parameters (especially mean and variance) may change according to the space-time circumstances.

Besides, in order to industrialize the use of these prediction models, some features must be adapted to ease the prediction process of new observations. Indeed, if we want to predict the number of bodily injury crashes that may occur on an unknown road, we must transform the model to include exposure variables as *offsets* in the model. Knowing the average traffic flow recorded on this road, as well as its length, we will thus set a baseline risk score that will then change according to space-time features selected.

Finally, telematics data processing and enrichment are still problematic for the moment, since it may take a long time to be achieved. Nevertheless, machine learning and big data algorithms that are rapidly developing, will soon ease this type of analyses. One of the potential solution to industrialize telematics risk studies is to count on real-time data processing. Indeed, telematics recordings are meant to feed a constant growing data flow, and enable in the end a real-time motor pricing. *Streaming* techniques and associated on-line machine learning algorithms such as *Vowpal Wabbit* [8] are dedicated to such real-time statistical analyses. *Vowpal Wabbit* should be moreover considered as an adapted tool to deal with telematics data, as its algorithm is very efficient to process high volume of data rapidly and to deal with text data (refer for instance to the keys constructed to determine road segments).

5.5.2 A project dedicated to road safety enhancement

Figure 5.7: Slogan of the European Road Safety Charter campaign

This research project was first designed to actively contribute to road safety enhancement. The basic idea can be summed up by the following sentences : if one can accurately qualify the inherent risk bore by each road of a network, drivers can be advised to travel on **less risky alternative roads** than the ones they intended to drive on while planning their route, thus reducing the own driver risk taking. If this process is expanded on a large scale, it may even reduce the overall bodily injury crashes count, especially if complementary associated to prevention actions promoting safe behaviour such as no-alcohol and no-drug driving. Such research projects must be seen by insurance companies as a way to be recognized by the society as active stakeholders of road safety enhancement.

In order to get an order of magnitude of how influential our risk models might be when used in a road safety framework, we calculated a proxy of the **number of lives that would have been saved** in Switzerland if such advice had been provided to road users.

To compute it, we used the datasets built on telematics data used to fit the models (both the train and test

[8] More details about it can be found on the following web page : https://github.com/JohnLangford/vowpal_wabbit/wiki

datasets were used). We associated first associated to each road its nearest neighbouring road, using a 1-nearest-neighbour algorithm on latitude and longitude data. This nearest neighbouring road is considered as the "alternative" road the driver would have taken in place of the one he drove on [9]. Then the risk scores of each initial road and each alternative road associated were computed, that is the predicted count of bodily injury crashes that may have occur on each road, thanks to the GBM fitted in the previous sections. Our goal was to calculate the risk spread Δ between the initial recommended road and its alternative. Thus, if the risk scores were equal or the alternative road riskier than the initial one, meaning that there is no switch, Δ was set to zero, otherwise, Δ was set to the spread value. The range of Δ computed were then summed up and divided by the sum of initial risk scores. Finally, this ratio was multiplied by the number of Swiss drivers who died in a car crash in 2013, that is 269 people.

In the end, we estimated that if such alternative roads had been proposed to Swiss drivers in 2013, **between 23 (using the count model fitted on roads proven risky) and 30 (using the count model fitted on significant roads) lives would have been saved**.

These calculations may nevertheless suffer from some limitations. When redirecting drivers on alternative roads, it implies a change in traffic flow on them, thus a potential change in risk. A trade-off has to be considered between the *redirected traffic* induced by this advice, that would increase the exposure to risk on the alternative road, and the initial risk qualification of the alternative road.

Finally, if more features dealing with road quality (road surface quality, measures undertaken to cope with wide puddles, ...) are further included in the risk prediction model, this type of risk analyses can be also beneficial to city regulators. Indeed, they may help in specifying "black-spots" analyses (see section 2.4.3) to a road level, and accurately target road maintenance investments.

5.5.3 Converting this research study into a business product

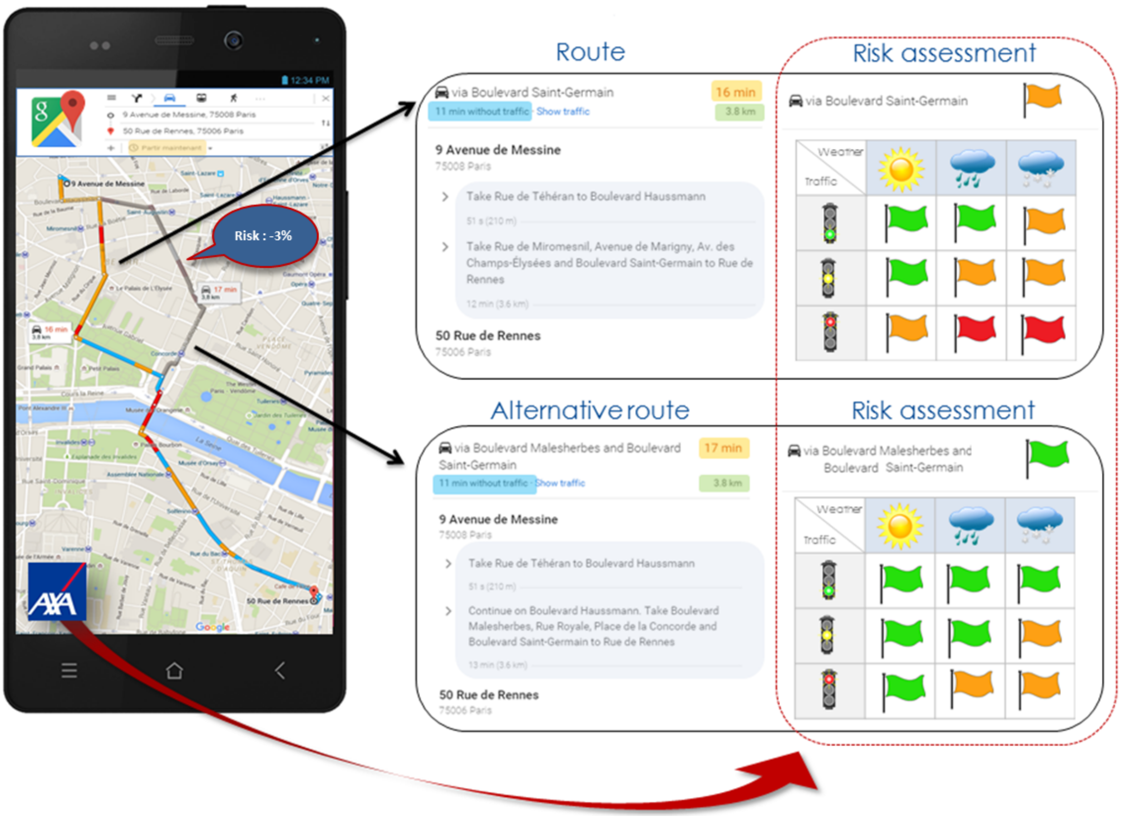

Risk : -3%
Route
via Boulevard Saint-Germain
16 min
11 min without traffic · Show traffic
3.8 km
9 Avenue de Messine
75008 Paris
Take Rue de Téhéran to Boulevard Haussmann
51 s (210 m)
Take Rue de Miromesnil, Avenue de Marigny, Av. des Champs-Élysées and Boulevard Saint-Germain to Rue de Rennes
12 min (3.6 km)
50 Rue de Rennes
75006 Paris
Risk assessment
via Boulevard Saint-Germain
Weather
Traffic
Alternative route
via Boulevard Malesherbes and Boulevard Saint-Germain
17 min
11 min without traffic · Show traffic
3.8 km
9 Avenue de Messine
75008 Paris
Take Rue de Téhéran to Boulevard Haussmann
51 s (210 m)
Continue on Boulevard Haussmann. Take Boulevard Malesherbes, Rue Royale, Place de la Concorde and Boulevard Saint-Germain to Rue de Rennes
13 min (3.6 km)
50 Rue de Rennes
75006 Paris
Risk assessment
via Boulevard Malesherbes and Boulevard Saint-Germain
Weather
Traffic

Figure 5.8: Insurers potential added value to routing services

[9] This approximation was done to ease the calculation. Indeed, the most nearest neighbouring road may not always lead to the same end point as the initial road or may not always be directly linked to the road at stake.

Warning drivers about the risk they take while travelling on a road in place of another according the space-time circumstances they face cannot be done without creating a proper business product directly dedicated to retail motor insurance customers. Services such as the alternative road proposal must be easily available for any driver who is planning his route. One of the most convenient way to do it would be to create an API (Application Programming Interface) linked to a routing service provider such as Google Maps.

Figure 5.8 gives an idea of how it might work. The drivers will first plan their routes thanks to a routing service provider on their smart-phones for instance. The routing API will give him back two possible routes to go from one point to the other, with a range of specific parameters. In figure 5.8 for instance, the routing was provided by Google Maps, that derives also a travelling duration according both to the traffic flow intensity it forecasts and to a normal traffic situation (see blue and yellow frames), as well as route instructions. On top of that, an insurance company API will retrieve these pieces of information, enrich them with environment and weather data especially, and use them to create a risk score for each proposed route thanks to a machine learning model such as the ones tested in this study. According to the risk scores computed and their own risk appetite, they will select the route they favour.

In the end, this type of customer-oriented product may both to improve road safety, and thus potentially decrease claims financial charges for insurers. Furthermore, it is very likely to position the insurance industry as an active sponsor of road safety and improve the way insurers are perceived by the public.

Conclusion

The goal of this report was to change the way regulators and actuaries usually look at car crash risks. Instead of focusing on behavioural and individual characteristics of drivers, we decided to study car crash risk from both a temporal and geographical point of view. The main purpose was to prove that motor risk may arise directly from specific space-time circumstances.

In this conclusion, we will synthesize the main results derived from the statistical studies carried out and give some suggestions of further developments and applications of such analyses.

Daily claims count can be explained to a large extent by weather and time features

Daily claims count was examined both as a time series, that depends on time and seasonal features, and as a count record that may be impacted by external phenomena such as traffic flow intensity or weather circumstances. It enabled us to retrieve some relevant time and weather features to model and predict the daily temporal evolution of motor collision claims count :

- **Traffic flow intensity** was proven to be an exposure variable : the more vehicles are travelling on the roads, the automatically greater is the number of collision claims recorded.
- The proportion of daily collision claims per driver decreases by 25% on **Sundays**, and by nearly 19% on public holidays.
- The number of daily collision claims per driver raises by more that 27% on **snowy days**.

Nevertheless, these effects may greatly change from one region, locality or even road to another, according to their geographical specificities. Thus, claims count analyses must not only be carried out at a country level, but also at the local level, and especially at the road-level.

Telematics data can bring a strong added value to motor risk analyses if they are properly enriched by external data

Telematics data gather both space and time features as they real-time record GPS trip positions. They thus may be very useful to carry out space-time motor risk analyses. Nevertheless, their collection process is still in the early stages, and the data recorded cannot be directly included in statistical models as things stand. They must be properly processed and enriched using external data to actually bring value to statistical analyses.

As telematics products are still in a pilot phase, data processing techniques are not properly stated not industrialized. When using these datasets, actuaries must be aware that this challenging stage may be time-consuming. We thus designed our own data processing method which goes through the following steps :

- Trips data had to be completed using a **routing** algorithm that derives the existing roads of the network on which drivers travelled, and retrieve the traffic that went through them.
- Claims data had to be **reverse geocoded** in order to attached a claims count to each road of the network under scrutiny.
- Trips and claims data had to be **merged** in order to link, for each road, its exposure parameter with its risk parameter.
- **External open data** (weather, topography and environment data) were finally used to associate to each road its specific surroundings and topographic characteristics, as well as weather features.

Each road is inherently risky, whatever the driver behaviour

Our goal was to prove that car crash risk can be partly directly put down to roads inherent risk factors, and not only to risky driver behaviours. We thus gradually included topographic, weather and surroundings features in our statistical models. At each step, the Gini index rose, proving that these features are relevant to characterize car crash risk on a road level.

We showed in the end that weather features are less influential than surroundings characteristics to explain roads inherent risk magnitude. Among the most influential features are : the traffic recorded on that specific road, the length of the road and its type (highway, national roads or smaller one).

The inherent riskiness of roads varies according to their specific geographical features

Roads inherent risk can vary widely from one road to another according to their geographical characteristics and to the intensity of the traffic flow recorded on them. Here is a sum up of the main effects captures by our roads inherent risk scoring model :

- When the telematics **traffic count** exceeds roughly 400 cars, doubling the traffic count on the road increases the number of bodily injury crashes located on it by nearly 150%.
- When looking at roads of more than 1 kilometre **length**, the longer they are, the more numerous bodily injury claims count are recorded on them.
- Roads located in **high mountainous localities** (beyond 3500 meters high) count twice as many bodily injury crashes as the other roads.
- **Primary roads**, that is mostly highways, record twice as many crashes as tertiary roads, that is roads linking medium-sized towns, and smaller types of roads.

How can these risk studies be used and improved in the future ?

The analyses performed on telematics data in this report are still at the draft stage. Indeed, their collection process and external enrichment still need to be improved. It requires nevertheless more accurate databases and claims telematics data recorded on longer periods, but we have no doubt that it will become possible in the coming years thanks to the development of such products.

In the future, risk assessments using telematics data may be beneficial to insurance companies from at least two point of view. On the one hand, risk analyses based on telematics behavioural features will obviously be exploited to refine motor pricing techniques and improve portfolio segmentations. On the other hand, space-time risk analyses such as the ones carried out in this report may strengthen the influence of insurance companies as road safety enhancement stakeholders. Indeed, the outcomes of such space-time motor risk assessments may for instance feed a driving assistance mobile application displaying warning messages for each road of a planned route, and proposing alternative less risky roads to drivers, according to the weather and traffic conditions they face. Moreover, such analyses may also be beneficial to regulators : by pointing out roads that are inherently risky, insurers can help governments to better allocate their road maintenance budgets and give priority to so-called "black-spots" locations.

Bibliography

[1] European commission statistics service on road safety. http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm.

[2] Swiss federal statistical online platform displaying geo-located data. https://map.geo.admin.ch/?X=190000.00&Y=660000.00&zoom=1&topic=ech&lang=en&bgLayer=ch.swisstopo.pixelkarte-farbe&catalogNodes=457,687,702&layers=ch.astra.unfaelle-personenschaeden_alle.

[3] M. A. Abdel-Aty & A. E. Radwan. Modelling traffic accident occurrence and involvment. *Accident Analysis and Prevention*, 32:633–642, 2000.

[4] M.M. Ahmed & M.A. Abdel-Aty & H. Huang & B. Guevara. Exploring a bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis and Prevention*, 43:1581–1589, 2011.

[5] M. Bateson & D. Nettle & G. Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology letters*, 2(3), 2006.

[6] J. Bentham. *The Panopticon*, volume 4. 1780.

[7] Cognizant. The new auto insurance ecosystem : Telematics, mobility and the connected car. Technical report, August 2012. (Cognizant est une entreprise américaine de conseil en informatique).

[8] B. De Laet. Regression trees and ensemble of trees in P&C pricing. *Master thesis written at AXA Belgium*, 2013.

[9] O. De Mouzon & N-E. Faouzi & B. Nowotny & J-M. Morin & E. Chung. Data fusion for traffic and safety indicators : the intelligent roads perspectives. *Proceedings of the 13th World Congress on Intelligent Transport Systems (ITS)*, October 2006.

[10] C. Doz. Cours de séries temporelles linéaires, note sur les tests unitaire. *ENSAE 2nd year courses*, 2004. http://lacote.ensae.net/SE206/Cours/Note%20sur%20les%20racines%20unitaires.pdf.

[11] J.H. Friedman. Stochastic gradient boosting. *Department of Statistics Standford University*, 1999.

[12] J.H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

[13] T.F. Golob & W.W. Recker. Relationships among urban freeway accidents, traffic; flow, weather; and lighting conditions. *Journal of Transportation Engineering*, August 2003.

[14] J.E. Gordon & J.J. Gibson. The epidemiology of accidents. *American Journal of Public Health*, 1949.

[15] W. Jr Haddon. Advances in the epidemiology of injuries as a basis for public policy. *Public Health Report*, 95:411–421, 1980.

[16] P. Händel & I. Skog & J.Wahlström & F. Bonawiede & R. Welch & J. Ohlsson & M. Ohlsson. Insurance telematics : opportunities and challenges with the smartphone solution. *IEEE Intelligent Transportation Systems Magazine*, July 2014.

[17] O. Lopez. Insurance econometrics. *ENSAE Actuarial Science program courses*, 2014.

[18] OECD & International Transport Forum & International Traffic Safety Data and Analysis Group. Road safety annual report. 2014.

[19] M-H. Pham & A-G. Dumont. Fusion of risk indicators aiming to predict near future traffic crash risks on motorways. *11th Swiss Transport Research Conference*, May 2011.

[20] M-H. Pham & E. Chung & O. De Mouzon & A-G. Dumont. Season effect of traffic : a case study in switzerland. *Seisan Kenkyu Production Research*, 59(3):214–216, 2007.

[21] M-H. PHAM & O. DE MOUZON & E. CHUNG & N-E. FAOUZI. Sensitivity of road safety indicators in normal and rash cases. *10th International Conference on Application of Advanced Technologies in Transportation*, May 2008.

[22] R. TIBSHIRANI & T. HASTIE & J. FRIEDMAN. *an Introduction to Statistical Learning*. Springer Texts in Statistics, 2009. http://statweb.stanford.edu/~tibs/ElemStatLearn/.

[23] U. TINGUELY. Points noirs sur les routes nationales 2011-2013. *Rapport de l'Office Fédéral des Routes (OFROU), Département fédéral de l'environnement, des transports, de l'énergie et de la communication DETEC*, July 2014.

[24] WORLD HEALTH ORGANIZATION. World report on road traffic injury prevention. 2004.

[25] D. YOON & J. CHOI & H. KIM & J.KIM. Future automotive insurance system based on telematics technology. *10th International Conference on Advanced Communication Technology, IEEE*, 1:679–681, February 2008.

List of Tables

List of Figures

Appendices

APPENDIX A

Road safety handling in Switzerland

A.1 Bodily injury crashes locations in Switzerland

This map overlaps a proxy of the daily average traffic per region in 2014 (assuming it was nearly the same from 2011 to 2013) and bodily injury car crashes locations reported by the Federal Roads Office between 2011 and 2013.

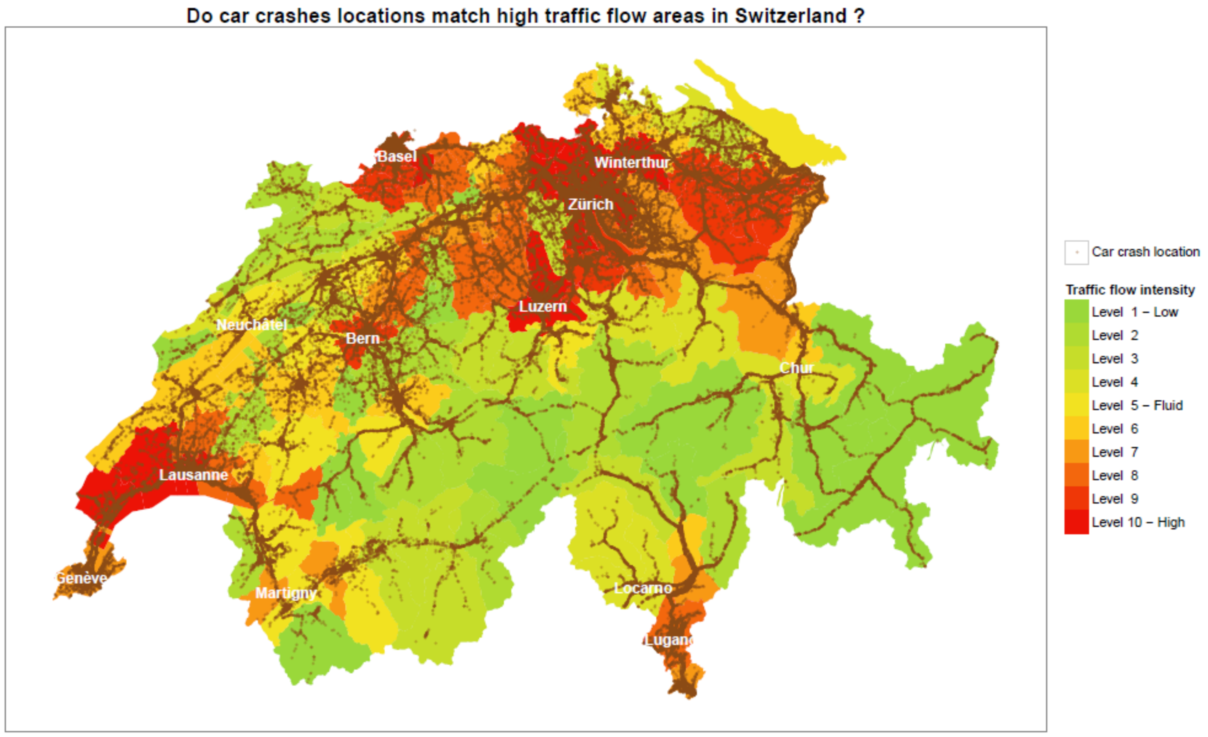


Figure A.1: Bodily injury crashes locations over traffic density

A.2 "Black spot" definition

Car crashes are associated to a "black-spot" location as soon as their characteristics (severity, road type) make their "norm" exceed a certain threshold. The "norm" used is a function of the number of severe crashes and slight injury crashes counted in a certain radius around a specific road, and overweights severe crashes :

$$2 \times U_{Severeinjuries+Fatalities} + 1 \times U_{Slightinjuries}$$

where : $U_{Severeinjuries+Fatalities}$ stands for the number of car crashes that lead to severe injuries or deaths and $U_{Slightinjuries}$ stands for the number of car crashes that lead to slight injuries.

Road type	Radius around the point (meters)	"Norm" threshold
Motorway	250 m	$\geqslant 8$
Countryside	150 m	$\geqslant 5$
Urban area	50 m	$\geqslant 5$

Table A.1: "Black spot" definition and threshold according to the type of road [23]

A.3 Traffic counters

Switzerland possess a wide network of road traffic counters which are mapped on the figure A.2, call the Swiss Automatic Road Traffic Counts (SARTC). They are located on federal highways and count the daily traffic. Aggregated databases per month are available on the FEDRO website : http://www.astra.admin.ch/verkehrsdaten/00299/00301/index.html?lang=en.

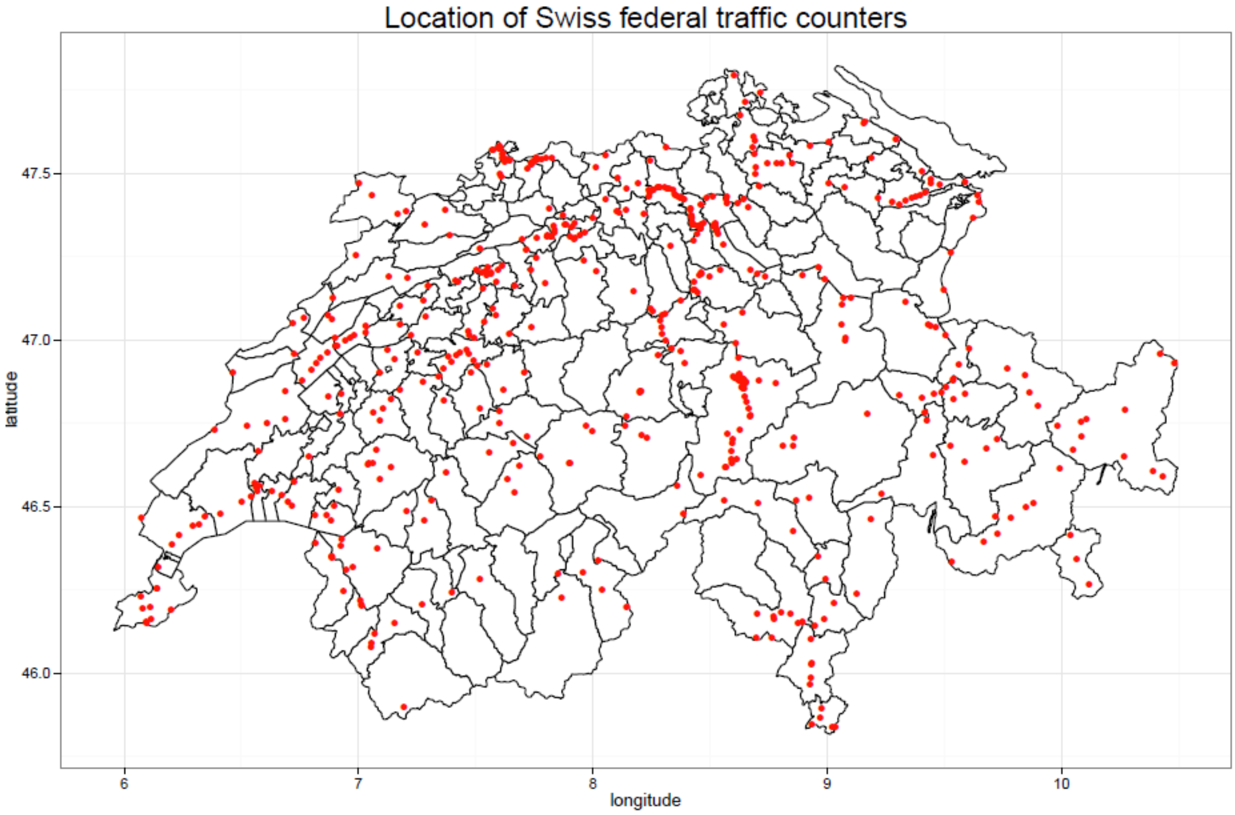


Figure A.2: Location of FEDRO automatic traffic counters

APPENDIX B

Claims count temporal analysis

B.1 Time series analysis

B.1.1 Schmidt-Philips unit root test

Schmidt-Phillips test : As the Augmented Dickey-Fuller test results may not always suitable to trended time series [10], the Schmidt-Phillips makes the null hypothesis clearer. Schmidt-Phillips test assumes the y_t times series writes $y_t = \alpha + \beta \times t + u_t$, with $u_t = \rho \times u_{t-1} + \epsilon_t$ [1] being non-stationary under H_0 (i.e. $|\rho|$=1) and stationary under the alternative hypothesis H_1.

Coefficient	Estimate	p-value	Significance
α	78.94	< 2e-16	***
β	0.02	< 2e-16	***
ρ	0.10	7.73e-07	***

	Test-statistic	1% critical value
T_ρ	-4929.35	-25.2

Table B.1: Schmidt-Phillips unit root test results

B.2 Claims count model

B.2.1 Weather station locations

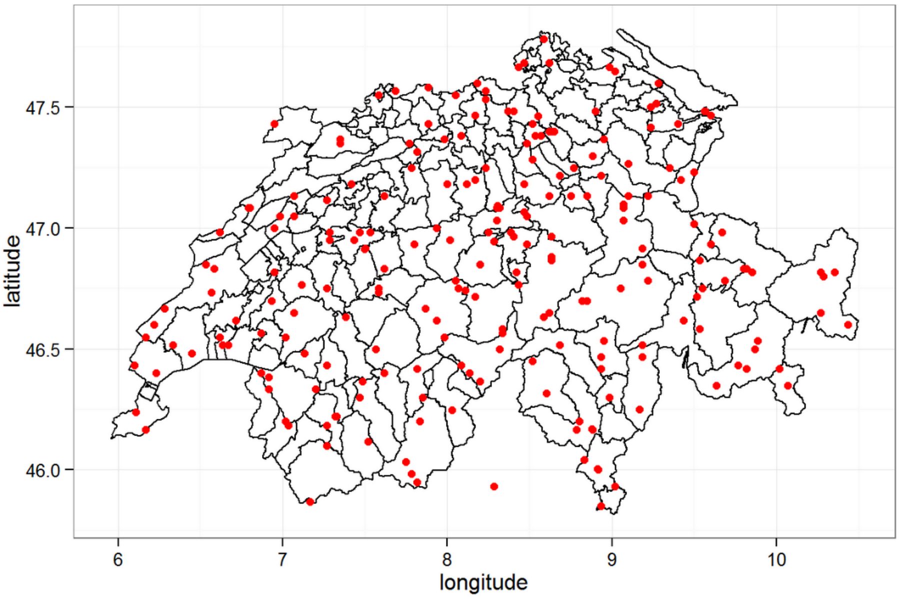


Figure B.1: NOAA weather stations locations

[1] ϵ_t is not assumed to be a white noise any more contrary to the Augmented Dickey-Fuller test.

B.2.2 Gradient Boosting model

Choosing the number of weak-learner trees fitted

Figure B.2 was used to determine the number of trees to grown in the GBM. Around 200 trees, the validation deviance decreasing slope clearly slows down, without entering the over-fitting stage. We did not choose more than 200 trees even if the valid devience minimum was not reached at that point, because the marginal gain in using more trees would have been mere. Several GBM were fitted with changing tree number to confirm this choice.

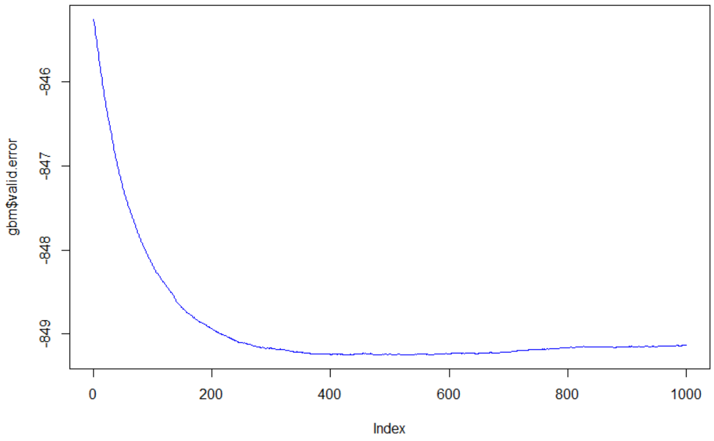


Figure B.2: Evolution of the validation deviance according to the number of trees grown in the GBM

An example of over-fitting while using a GBM

In order to show a situation in which a gradient boosting model is strongly over-fitting, we computed a gradient boosting model including the same offset (that is log(traffic x portfolio size)) and same input variables as what was presented section 3.4.2, and added traffic variable as an input variable. It means that the traffic variable was taken twice into account, favouring thus over-fitting. Figure B.3 shows the partial dependence chart derived from this model. Dependence trajectory is not smooth at all, and the trends spotted by the model are not easily interpretable and seem even counter-intuitive.

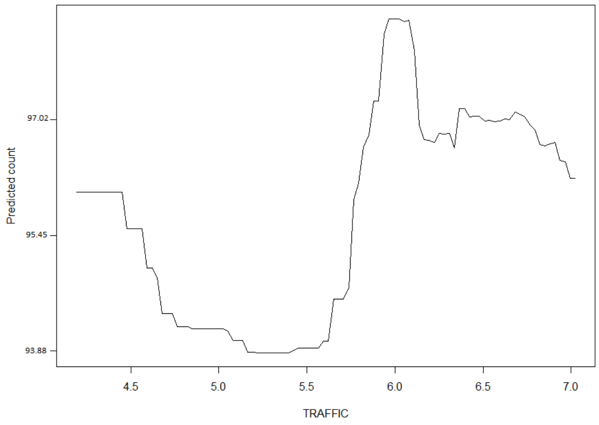


Figure B.3: Partial dependence plot of traffic flow over daily claims count

How are partial dependence plots computed ?

Explanations about the computation of partial dependence plots are derived from [8].

Partial dependence plots help to understand the dependence of the prediction $\hat{f}(x)$ on the joint values of the input variables. Consider the sub-vector $\mathbf{x}_l$ of size $l < p$ of the input variables $\mathbf{x} = (x_1, ..., x_p)$ indexed by $G \subset \{1, ..., p\}$. Let C be the complement of G, such that $G \cup C = \{1, ..., p\}$. We now determine the partial dependence of $\hat{f}(x)$ on the subset $\mathbf{x}_l$, using the following formula :

$$\hat{f}(\mathbf{x}_l) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\mathbf{x}_l, \mathbf{x}_{ic})$$

where $\{x_{1c}, ..., x_{nc}\}$ correspond to the values of the complementary subset $\mathbf{x}_c$ from the training set. This computation requires a prediction of the response variable for each set of joint values of $\mathbf{x}_l$. This is generally computationally intensive, except in the case of tree-based gradient boosting algorithm.

In this case, a "weighted traversal method", introduced by Friedman [12], can be used. This method computes the dependencies by using only the tree, and not the input data. At the root of the tree, we set a weight value of 1. We then go down past all the internal nodes. If the splitting variable is in the subset $\mathbf{x}_l$, we follow further the appropriate child and we do not modify the weight. On the contrary, when the splitting varaible is part of the complementary subset $\mathbf{x}_c$, we follow both childs of the internal node. The current weight is multiplied by the fraction of observations that belongs to the left and risk child respectively. If we reach a terminal node, we assign the current weight to that node. When all nodes were crossed, the value of $\hat{f}(\mathbf{x}_l)$ is computed as the weighted average over the terminal node responses, with the weights calculated as above. In the boosting framework, we jest take the average over all the trees.

For the purpose of visualization, it is better to limit the size of the subset G to one or two covariates. A size of one helps us to see the marginal dependence of the estimator to a single variable. It corresponds to the plots we analyse in this report. A set of two covariates gives an interesting insight in interactions between two variables. It corresponds to the so-called "interaction plots". Using a higher set size is not recommended, since in that case we cannot visualize it easily.

APPENDIX C

Roads risk scoring using telematics data

C.1 Roads classification : GLM comparison

Table C.1 introduces the estimation results of a logistic model fitted on the dummy variable that equals 1 if at least one bodily injury claim was registered on the road and 0 otherwise.

Covariate	β *(all other things remaining equal)*	**Significance**
Log(Traffic count)	- 7.16e-04	
Log(Minimum distance)	- 9.61e-02	***
Log(Maximum distance)	+ 1.74e-01	***
Speed spread	+ 2.02e-02	***
Maximum speed	- 5.05e-03	*
Maximum height	+ 4.82e-05	
Maximum slope	+ 2.61e-03	
Traffic fog	+1.35e-02	**
Traffic rain	- 1.46e-03	
Traffic snow	+ 4.34e-02	***
Traffic freeze	-3.83e-03	*
Primary road	+ 1.13	***
Secondary road	+1.63	***
Tertiary road	+ 1.09	***
"Turn" manoeuvre	+ 2.54e-01	***
"Junction" manoeuvre	- 1.441	***
"Roundabout" manoeuvre	+ 6.78e-01	***
Graubünden region	-1.10e-01	
Ticino region	1.04e-01	
Zürich region	-4.17e-01	***
Geneva region	5.82e-01	***
Null deviance	16 987 on 12 615 degrees of freedom	
Residual deviance	14 672 on 12 594 degrees of freedom	
AIC	14 176	

Table C.1: Roads classification Generalized Linear Model results

C.2 Bodily injury crashes count prediction : GLM comparison

Table C.2 introduces the estimation results of a Poisson GLM fitted on the bodily injury claims count of the roads spotted by telematics drivers trips.

Covariate	**Influence over bodily injury crashes count** *(all other things remaining equal)*	**Significance**
Log(Traffic count)	+ 3.15%	***
Log(Minimum distance)	- 11.30%	***
Log(Maximum distance)	+37.69%	***
Speed spread	+ 2.50%	***
Maximum speed	- 1.69%	***
Maximum height	+ 0.03 %	***
Maximum slope	- 0.67%	***
Traffic fog	+1.15%	***
Traffic rain	+ 0.94%	***
Traffic snow	+ 2.14%	***
Traffic freeze	- 0.43%	***
Primary road	+ 188.06 %	***
Secondary road	+ 175.39%	***
Tertiary road	+ 55.26%	***
"Turn" manoeuvre	-7.96%	***
"Junction" manoeuvre	+ 0.61%	
"Roundabout" manoeuvre	+ 30.97%	***
Graubünden region	+ 0.83%	
Ticino region	+ 21.41%	***
Zürich region	-28.82%	***
Geneva region	+ 25.35%	***
Null deviance	126 979 on 12 615 degrees of freedom	
Residual deviance	83 185 on 12 594 degrees of freedom	
AIC	99 172	

Table C.2: Crashes count prediction Generalized Linear Model results