NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Data Science for Connected Car Insurance

Use of Trips Raw Telematics Data for Knowledge Discovery and Customers Profiling

CONFIDENTIAL

Enrico Spada (M2016288)

Internship report presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DATA SCIENCE FOR CONNECTED CAR INSURANCE

by

Enrico Spada

Internship report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Advisor: Professor Pedro da Costa Brito Cabral

Internship Coordinator*:* Olivier Claeys (Chief Actuary Officer)

May 2018

DEDICATION

To my Grandfather Luigi who always inspire me to pursue any undertaking.

ACKNOWLEDGEMENTS

This thesis becomes a reality with the kind support and help of many individuals. I would like to extend my sincere gratitude to all of them.

Foremost, I want to offer this endeavour to my professor and supervisor Pedro Cabral for the guidance and support, fundamental to accomplish this research. Thank you for your kindness, helpfulness and always wise directions.

I am highly indebted to Sterling Insurance for the trust bestowed upon me. Thank you for the internship opportunity and for giving me the chance of developing my thesis on this extremely interesting and fascinating topic. I express a special thanks to the entire Actuarial Department in which I carried out my internship. Thank you Alessandra, Alessia, Giovanni, Laura, Lorenzo, Marco, Mattea, Olivier and Veronica.

I would like to express my most sincere gratitude to my internship coordinator, Olivier Claeys, and to Giovanni Di Carlo for imparting their expertise and knowledge in this study. Thank you for mentoring me throughout my first professional job experience and for instilling in me invaluable teachings.

I would like to convey my gratitude towards my wider Family for the serenity and the love. To my beloved and supportive *namorada*, Luisa who is always by my side when times I needed her most and helped me a lot in the making of this study.

Much gratitude goes to Martinho Silvestre for sharing his deep knowledge, his incredibly sound, analytical approach in solving any problem, and his dedication toward understanding. Thank you for your guidance allowing me to build fundamental skills for this study and for my professional career.

Finally, I thank Portugal, Nova IMS, all the amazing professors and colleagues I met and helped me out with their abilities and wisdom.

DATA SCIENCE FOR CONNECTED CAR INSURANCE

ABSTRACT

This report presents all data science processes designed and implemented during the internship at the Actuarial Department of Sterling Insurance[1] (Italy). The project developed a complete data science solution, organized according to Cross-Industry Standard Process for Data Mining. The objective is to study in-depth – for the very first time – trips raw telematics data, and to discover actionable knowledge that can be applied to generate value for the business.

The research is based on trips raw telematics data generated over 5 months by telematics black-box devices installed in the cars of 937 customers. The data are solely related to trips, with granularity at the finest level of individual geospatial coordinate sets composing trajectories. The features describing each timestamped GPS coordinate set are average speed in the last second, heading, GPS quality, meters travelled since previous position. The data sources consist of semi-structured data stored in several flat files in their native format, batch extracted from the data lake.

Starting from trips raw telematics data at the granular level of geospatial coordinate sets, they are extensively studied and enriched with additional open data sources exploiting spatial join operations. Next, a complex concatenation of data preparation tasks is performed to obtain the final dataset, aggregated at the granular level of trips and described by 117 features. The final dataset is fed to the k-means algorithm for discovering patterns over trips characteristics. Patterns are studied considering the overall portfolio, regardless of driver and intentionally neglecting historical or personal information.

The study concludes by deploying the clustering results to profile customers, bringing to a new level the risk knowledge of the line of business about its customers. This discovery opens a world of new possibilities, some of the uncountable examples are improve pricing, using results in fraud detection and offering new services and overall risk prevention for customers.

KEYWORDS

Data Science; Car Insurance; Raw Telematics Data; Clustering; Risk Knowledge

[1] This is a fictitious name to guarantees confidentiality.

INDEX OF THE TEXT

INDEX OF FIGURES

INDEX OF FIGURES

GLOSSARY

Actuarial Science	Statistics relating to insurance, estimating loss reserves and developing premium rates.
ANIA	Italian Association of Insurance Companies
Claim	Demand by an individual or corporation to recover, under a policy of insurance, for loss that may come within that policy.
Insurance Company	Company authorized to sell insurance to the general public, in compliance with the Community guidelines on direct insurance
Insurance Policy	Written contract of insurance between the insurer and the policyholder delineating the coverage term, the insurance policy limits, the grant of coverage, exclusions and other limitations of coverage, and the duties and responsibilities of the insured in the event of a loss
MTPL	Motor third-party liability, which refers to a person's legal liability for the bodily injury and/or property damage sustained by another as the result of a motor vehicle-related accident.
Non-Life	Interchangeably used with "property and casualty" for describing insurance coverages other than life
Tariff	Refers to rates and coverages set and published by the rating bureau having jurisdiction. The rating bureau may be controlled either by an association of companies or by a foreign government.
Telematics Black-Box	Device installed in customer's vehicles to gather and transmit data related to GPS coordinates, direction and speed
Telematics Data	Data generated by a telematics black-box installed on customers' vehicles.
Underwriting	Process of determining whether to accept a risk and, if so, what amount of insurance the company will write on the acceptable risk, and at what rate.
Voucher	Anonymized identifier associated to customers of telematics products, allowing to univocally connect their black box to Sterling Insurance's CRM system.

1. INTRODUCTION

Telematics is a combination of the words telecommunications and informatics. It was first coined in 1978[2] broadly referring to any transfer of information over telecommunication systems. Today, it is generally referred to its best application: vehicles. Telematics applied to motor vehicles allows remote monitoring of vehicle's locations and movements. The technology is widely adopted for monitoring fleets of vehicles from courier companies to emergency services.

One of the most successful adoption of Telematics is in insurance business. These technologies can affect every line of business, but the current project focuses solely on car insurance. Italy is pioneering this field and accounts for the highest penetration of insurance products bound to telematics devices. Telematics data represents an outstanding opportunity for insurance companies because it has made possible to:

- improve risk selection
- innovate pricing
- establish valuable relationship with customers enhancing loyalty
- shift from a passive role toward a proactive role in the claims settlement process
- offer new services
- create Business-to-Business opportunities

Telematics has the potential of significantly change the car insurance business because it affects all value chain and creates tangible added value. For Sterling Insurance, the real disruptive potential of vehicle telematics is related to new, high value services and analytics[3]. To unlock this potential, Sterling Insurance needs to develop capabilities and processes to implement raw telematics data into new routines and operations.

This project is relevant for Sterling Insurance because it studies for the very first time raw telematics data and builds valuable internal know-how about methods and processes required to extract actionable knowledge. The internship is integral part of Sterling Insurance's industrial plan, which consists of expanding its share of telematics portfolio and start developing new, value added services based on raw telematics data to offer both customers and other companies. In order to offer new services, it can be sufficient to define drivers' behaviour or to detect unusual routes.

For insurance companies, the project is relevant because reassure themselves about the quality of information provided by telematics black-boxes and how to interpret it. Additionally, this is an extremely important topics for Research & Development in car insurance business because promises outstanding opportunities in terms of competitive edge and value creation.

[2] L'informatisation de la societe: Rapport a M. le President de la Republique, 1978

[3] For example, this technology makes available data becoming the base for analysis and accurate reconstruction of crash dynamics.

This project is innovative because makes use for the very first time of trips raw telematics data. Raw telematics data are jealously safeguarded from competitors and Sterling Insurance took possession for the first time of raw telematics data on December 2017. Therefore, this research is advancing the state of the art in the context of vehicle telematics applied to car insurance business.

Currently, Sterling Insurance is replicating the methodology defined in this research to study crash raw telematics data. The next project will most likely consist of discovering patterns over the combination of trips and crash raw telematics data. Based on those information, it will be possible in the future to develop real-time predictive model to prevent drivers from crashing.

This project performed a targeted study with high added-value by in-depth analysing for the first time raw telematics data and offering a possible business application of knowledge discovered. The study offered a valuable understanding of raw telematics data which is at the hearth of the relevance of any new service.

This study aims to discover whether trips raw telematics data contain useful information. The research objectives are to design and implement data science methods and processes required to analyse and extract knowledge starting from trips raw telematics data, in order to discover new perspectives of risk knowledge associated to customers behaviour. Specifically, the objectives of this study were to:

1. Define a clear data mining problem starting from the understanding of business objectives;
2. Explore trips raw telematics data to develop understanding and assess data quality;
3. Integrate raw data with additional data sources freely available in order to enrich its context;
4. Design all those data preparation tasks required to obtain the final dataset to feed the clustering algorithm, starting from the initial raw data;
5. Interpret clustering results and deploy the new knowledge to offer a new perspective over risk insured and suggest possible actions to take.

1.1. LITERATURE REVIEW

The following literature review details relevant contributions on the methodologies adopted to carry out this research. In the first part it is discussed the lack of scientific literature available on raw telematics data. In the second part is illustrated the scientific literature related to Data Mining methodologies and practices implemented to extract actionable knowledge from raw telematics data. It starts off by presenting two of the main approaches to Data Mining, which were introduced by two developers of analytics software. Next, the literature related to common tasks solved in the data mining process. Special consideration is given to scientific works on clustering. Then, it presents literature related to most valuable *R packages* adopted in this research. To conclude, for completeness of the study it mentions a brief overview on trajectory data mining literature.

Discussion on vehicle telematics literature

Considering the innovative research of this study, there are not many contributions available. Most of the knowledge on this innovative field is internally developed by companies proprietary of the raw data and safely shielded from competitors. In fact, this report is confidential and will not be published. On the opposite, extensive literature is available for traditional vehicle telematics related to data at the granular level of monthly or daily summary statistics.

The only studies partially resembling the research field of this project are based on data made available by AXA exclusively to participants of a Kaggle competition in 2015[4]. The participants studied spatial trajectories generated by telematics device installed on 2 736 vehicles; each driver was associated to 200 trips. The data solely consisted of series of chronologically ordered geospatial coordinates and timestamp at the sampling rate of 1 Hz[5]. Huang & Nikulin (2016) published a scientific paper based on their experience in the Kaggle competition and they developed an unsupervised methodology enriched with substantial feature engineering to recognize the driver fingerprint and discover abnormalities in trajectories. Another study based on the same data presented an alternative consisted of clustering over heatmaps representing driving style displayed as average speed on the x-axis and the corresponding acceleration/braking pattern on the y-axis; multiple heatmaps were generated by bucketing the speed (Wüthrich, 2016).

The current study substantially diverges from the above for three main reasons. First, it is in the context of Research & Development in an insurance company pioneering vehicle telematics. Second, the data sampling rate is much lower; one of the implications of this aspect is that it is extremely difficult to characterize the trajectory of short trips for which very little information is available. third, the data are not anonymized – with exception of customers personal information – making it possible to enrich them with additional open data sources.

Data Mining literature

Data Mining is described as the process of discovering knowledge by searching through large amount of data (Gorunescu, 2011). It can be considered as an iterative and cyclical process that requires goals and objective to be specified (Shearer, 2000). Data Mining encompasses a series of pattern recognition technologies and mathematical and statistical techniques to discover correlations,

[4] The page of the competition can be found at: https://www.kaggle.com/c/axa-driver-telematics-analysis

[5] One position was recorded and transmitted by the telematics black-box every second.

patterns, trends or relationships hidden in the data structures (Heikki & Mannila, 1996). The final objective is "to summarize the data in novel ways that are both understandable and useful to the data owner" (Fayadd et al., 1996).

Data Mining involves a structured approach for its implementation (Han & Camber, 2001). There are several different data mining methodologies, but there is no standard one. Therefore, software vendors have designed approaches that are most suitable with the design of their own advanced analytics software solution. Two popular methodologies for developing a data mining project are: Sample-Explore-Modify-Model-Assess (SEMMA)[6] proposed by SAS Institute (SAS Enterprise Miner, 1999) and CRoss-industry Standard Process for Data Mining (CRISP-DM) proposed by SPSS (Chapman, 1999).

SEMMA (SAS Enterprise Miner Documentation) is specifically designed to work with the Enterprise Miner software. This data mining process offered by SAS articulates in five stages: sample, explore, modify, model and assess. This methodology differs from CRISP-DM because it does not consider the business context in which the project is developed. Also, it lacks design and implementation phases.

CRISP-DM (Shearer, 2000) breaks down the life cycle of a data mining project into 6 phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The phases are connected by important dependencies and the cyclical nature allows different iteration of the process to trigger new, more focused business question. The business understanding phase focuses on data mining problem definition and planning to achieve the objectives.

The business understanding phase can be identified in the process of externalization of tacit knowledge and in the process of combination of explicit knowledge (Nonaka, 1991). Dialogue was an effective tool to articulate Sterling Insurance employees' tacit knowledge and share it into the business understanding phase. Examples of explicit knowledge sources are ANIA's *Italian Insurance* annual reports and Swiss RE Institute's *sigma research* annual publications. Also, it was very valuable "Unveiling the full potential of telematics – an Italy case study" (Dang, 2017) which focused on how vehicle telematics can bring value to insurers and consumers in Italian market.

The Exploratory Data Analysis (Tukey, 1977) defines techniques and tools for data analysis, visualization approaches and objectives pursued to investigate and explore telematics data. Principles for functional (Tukey, 2001) and truthful (Cairo, 2016) data visualization have been implemented for all visualizations.

The objective of feature selection is to exclude irrelevant and redundant features from the training set for improving results of machine learning algorithm. The literature on this topic for supervised learning algorithms is very rich distinguishing them depending on the algorithm used in wrapper, embedded, filter, and hybrid methods (Singh et al., 2016). On the opposite, considering the subjective nature intrinsic to unsupervised learning, it is more difficult to define methodologies for features selection.

There are several techniques used in data mining for clustering tasks, but not all of them can be applied to all situations (Estivill-Castro, 2002). Cluster analysis involves separating sets of data into

[6] Documentation can be found at:
https://support.sas.com/documentation/cdl/en/emcs/66392/PDF/default/emcs.pdf

groups that include a series of consistent pattern (Everitt, 2011). The most common techniques applied in this task are unsupervised learning algorithms and data visualization. Xu (2005) presented an extensive survey of clustering algorithms including the four basic steps of cluster analysis: feature selection, Clustering Selection, Cluster Validation and Results Analysis. In Survey of Clustering Data Mining Techniques (Berkhin, 2006), K-Means methods (Hartigan 1975; Hartigan & Wong 1979) are defined as the most popular clustering tool used in industrial applications. They are simple, straightforward and based on analysis of variance.

Clusters Profiling (Fawcett, 2013) and Data Visualization (Tufte, 2001) are further fundamental data mining tasks because allow to effectively communicate the results and insights discovered in order to impact on the business (Knaflic, 2015).

2. INSURANCE BUSINESS CONTEXT

The goal of this chapter is to describe the functioning of the insurance business and what vehicle telematics represents for this industry. This knowledge contributes to explain the relevance of this work and the reasons why this project is important to Sterling Insurance.

In the first section is described the insurance business of private companies from juridical and business economics perspectives. The implications of specific regulations are discussed, and the insurance business model is presented. Then, are explained some of the main technical indicators measuring performance and driving strategy objectives and tactical initiatives. To conclude, it is presented an overview of the Italian insurance industry focusing on the motor insurance market.

The second section extensively studies evolution of Telematics since its first appearance in 2003 as a small pilot test of Unipol Insurances aiming at understanding its applicability in road safety. The section concludes focusing on the reasons why companies should embrace this innovation. In fact, insurance industry is approaching a revolution triggered by the technological innovation in vehicles automated driving systems. This may reflect in a physiologic de-risking trend of risk profiles related to cars. The impact on car insurance business is an escalating decline of premiums. Car insurance companies need to find new sources of revenue and Telematics is one of the possible answers.

The last section presents Sterling Insurance as a connected car insurance because it is innovating to embrace this data-driven revolution by fostering spread of insurance products bound to a telematics device always connected to the company. The motor line of business of Sterling Insurance is analysed and corporate strategy is described in relation to its "Internet of Things and Telematics" strategy. To conclude, it is presented an overview of the internship.

2.1. THE INSURANCE BUSINESS

Insurance companies plays an economic and social role avoiding possible negative events. Considering the "risk-adverse" nature of clients, insurance companies represent the most immediate, rational and practical answer to the need of safety throughout time and events.

In Italian law, the *insurance* is a contractual relationship that exists when one party – the insurer – for a consideration – the premium – is liable to another party – the insured. According to article 1882 and 1917 of Italian Civil Code, the insurer is liable for: (1) recouping the damage or loss suffered by the insured in the event of claim; (2) indemnifying and hold harmless the insured from third-party liability; (3) compensating by capital or annuity in the case of human life and health events. The insurance business is regulated by the Code of Private Insurance. Article 1 defines the insurance business as the taking up and management of risks through the *contract of insurance undertaking*. Article 11 provides that insurance companies are distinguished in *life* or *property and casualty*[7].

Property and casualty insurances encompass all contracts regulated as "the insurer liability to recoup insured's patrimonial losses caused by a damaging event". Life insurances encompass all contracts regulated as "whole and term life insurance policies".

[7] Referred to interchangeably as "non-life".

The corporate purposes[8] of insurance companies must exclusively pursue life business or exclusively pursue property and casualty business. By derogation to this principle, insurance companies can be authorized to jointly operates in life and non-life businesses, but under obligation of separating the management as two different corporate purposes.

For this reason, insurance companies must find subsidiary firms pursuing the specific corporate purposes of offering services not essential to the insurance contract defined in article 1882 and 1917 of Italian Civil Code. In June 2017, Sterling Insurance founded the company S-Evolution which corporate purposes are to create new, valuable services based on vehicle telematics to offer Business-to-Business and Business-to-Consumer markets.

The Code for Private Insurance Companies[9] lists the only possible insurance policies – which are referred to as *regulatory classes*. It is not relevant to dwell into each regulatory class for both insurance businesses. However, it is important the macro distinction within the non-life business in: Motors classes and Non-Motors classes. Motors classes encompasses two policies:

- Class 3 – Land Vehicle: covering from all damage to or loss of land vehicles
- Class 10 – Motor vehicle third-party liability (MTPL): protecting against all liability arising out of the use of motor vehicles operating on the land

Since 1969, insurance policies of class 10 are compulsory for all motor vehicles in order to cover the person responsible against all third-party liability as described in article 2054 of Italian Civil Code. In practice, it is referred to motor third-party liability (MTPL) insurance and it protects the vehicle against unexpected expenses causing harm to others' health or property. In 1994 the tariff was liberalized. The obligation of MTPL insurance was a great business opportunity allowing insurance companies to live off this new market. Nowadays, the side effect is that several companies exclusively depends on the car insurance business, which profitability traditionally follows a cycle of alternative positive and negative periods. The innovation in automotive industry is evolving car security technologies – i.e. semi-automated cars with driving assist systems – and it may diminish risk. This would reflect on reduction of average premiums and revenue for insurance companies focused on car insurance business.

Once the insurance business has been introduced, it is time to present it from a business economics perspective describing its peculiar business model.

2.1.1. The Business Model

Insurance companies sell to customers insurance policies related to the regulatory classes listed in the Code for Private Insurance Companies. They operate with a very specific business model. In general, business models describe the way companies create and capture value. The main peculiarity of insurance business model is reversion of the business cycle.

Usually, business cycles start with incurring costs for purchasing productive factors and end with obtaining revenue for selling goods and services. On the opposite, insurance companies first collect

[8] General objectives of a firm, as listed in its legal documents which gave birth to the corporation.
[9] Legislative Decree 209 of 7th September 2005.

premium income, and at a later time – only in the case insurance events occurred – they bear costs connected to the insurance coverage granted.

The reversed production cycle makes the management of insurance companies radically different from any other industrial and commercial companies. The main effects of this reversion are:

- Future uncertainty of costs;
- Pricing of premiums based on uncertain costs;
- Centrality of financial management for life business.

Costs connected to the insurance coverage are incurred only in the case of claim event. Random, uncertain expenses cannot be easily estimated. *An et Quantum Debeatur*[10] refers to uncertainty about when or whether an insurance event will occur and about how much it will amount to. For car insurance policies, the uncertainty is related to all three aspects.

Future claims will bear uncertain costs dependent on the amount of damage and on the probability associated with those insurance events. Insurance companies must define *a priori* pricing for policies to offer the market. The premium is a fair balance between risk perceived by customers and risk insured by the company. Solving this problem requires to estimate the uncertain costs. Pursuing this task require to depend on the role of actuaries and to rely on extensive databases that forms the statistical basis to carry out reliable probability estimates for correct risk management.

Claims related to life insurance policies incur long after collecting premiums. This means insurance companies need to manage efficiently large volumes of liquidity cover future expenses. The condition of equilibrium between claims cost and current assets is realized when financial performance of liquidity is maximized while at the same time respecting protection and constraints provided by the legislator. The limitations imposed on insurance companies provide protection to policyholders by guaranteeing consistency between inflows of return on investment and outflow of claims settlement expenses.

2.1.2. Key Performance Indicators

Key Performance Indicators (KPIs) are strictly related to the business and are domain of Business Intelligence. Business Intelligence objective is to store data from several sources into a data warehouse and produce extraction for informing decision makers. The purpose of KPIs is to measure and monitor performance.

To not dwell excessively in this section, it is appropriate to solely consider some of the several KPIs measuring the technical performance. Among them are presented three general-purpose indicators and four metrics to monitor each line of business. Such three general-purpose technical KPIs allows comparison between companies and analysis of company's performance over time. These are general-purpose and meaningful to measure the overall performance of the company as well as the performance of individual line of business. These metrics are:

[10] Latin phrase widely adopted in juridical lexicon.

- Expense Ratio is the percentage of premium to cover costs of acquiring, underwriting, and servicing insurance and reinsurance. These expenses include advertising, employee wages and commission for the insurance agents. It is important to point out this indicator is not a measure of ending profitability, instead, it is a precursor to find an insurance company technical profitability.

- Loss Ratio represents total claims expenses divided by total earned premiums. This indicator measures whether a company is paying more claims than the amount is collecting in premiums. If the indicator is above 1, the company is not collecting enough premiums to cover claims expenses. Acceptable values for Loss Ratio vary depending on the insurance policy class.

- Combined Ratio is the sum of Expense Ratio and Loss Ratio. When applied to a company's overall results, the combined ratio is also referred to as the composite, or statutory, ratio. Used in both insurance and reinsurance, a combined ratio below 100 percent is indicative of a profitable technical management.

Besides the three general-purpose indicators, there exists indicators that exclusively measure the performance of individual line of business. The following are four of the many indicators monitoring each single insurance policy class:

- Average Premium: gross earned premiums divided by number of insured for a policy class
- Volume: number of insurance policy issued for a policy class
- Claim Frequency: number of claims notified divided by number of insured for a policy class
- Average Claim Cost: gross claims expenditure divided by number of claims settled for a policy class

Car insurance policies last twelve months but can start any time throughout the year. Therefore, according to business economics principles, these indicators are computed on a *pro rata temporis*[11] basis for the accrual period of risk exposure.

The indicators presented in this section are important for driving business initiatives and measuring the technical results of strategic decisions. However, in the very end, the impact of corporate strategy is measured by stakeholders using global corporate finance indicators.

2.2. THE ITALIAN INSURANCE INDUSTRY

In 2016, Italy registered a ratio of total insurance premiums to gross domestic product (GDP) equal to 8.2%, decreasing compared to year 2015 (-9.0%)[12]. After three consecutive years of growth, the decline is caused by a contraction in Life premiums due to market drop in index-linked policies

[11] Proportional to the time allotted.
[12] Source: Swiss Re, Sigma n° 3/2017

registered throughout the year[13]. Italy ranks fourth in Europe and eight in the world for premium collection, accounting for 3.4% of worldwide insurance business volume[14].

The total amount of premiums acquired by 31st December 2016 was €134,206M (-8.7% compared to 2015). The breakdown of total premiums is 23.81% for property and casualty business (-0.2% compared to 2015) and 76.19% for Life business (-11.0% compared to 2015). Due to the sharp shortfall recorded in life premiums, the incidence of property and casualty business on total business volume increased from 21.8% of 2015 to 23.8% of 2016.

By 31st December 2016, insurance companies authorized to operate in Italian territory are 111. In the last ten years, the number of insurance companies has shrunk by 34% considering the 163 companies in 2007. There are 108 insurance companies with registered office in Italy; Sterling Insurance headquarters is in Rome. Among them 12 operates in both life and non-life businesses, 55 operates exclusively in non-life business and 41 operates exclusively in life business. All the foreign insurance companies operate exclusively in life business.

The level of market concentration is quite high. Considering non-life business, the first five companies have a market share of 69.2% and the first ten companies have a market share of 83.3%. Regarding life business, the first five companies have a market share of 59.8% and the first ten companies have a market share of 73.3%. This trend is steady, however, on individual basis the market share may vary due to merge or transfers of customers portfolio.

Since the core business of Sterling Insurance is motor insurance, it is important to dig deeper into the motor insurance market.

2.2.1. Motor Insurance Market

Motor insurance business encompasses policies related to Class 3 – Land vehicles and Class 10 – Motor vehicle liability. The number of vehicle insured in Italy is 38.7 million estimating 2.9 million[15] uninsured. The motor insurance business accounts to 50.6% of total non-life business volume. Compared to previous year, Class 10 policies (Motor vehicle liability) decreased by 5.6% accounting for 42.3% of non-life premiums. The downturn resulted from the fierce competition between companies, which positive underwriting results allowed them to pursue marketing strategies of competitive pricing[16] Accounting for 8.2% of non-life business volume, Class 3 policies (Land vehicles) increased by 6.5% compared to year 2015 confirming the positive trend of the last two years.

Analysing technical indicators measuring the performance of motor insurance market, the combined ratio[17] was 97.6% in 2016 compared to 93.6% of previous year. This deterioration is caused by drop of average premium for Class 10 policies (Motor vehicle liability). Considering average premium is the

[13] Starting in March 2015, European Central Bank adopted the monetary policy of quantitative easing. resulting in a sharp drop of bank interests rate. This impaired financial performance of some life insurance policies making them unprofitable.

[14] Source: ANIA, Italian Insurance in Figures 2016

[15] Source: ANIA

Source: ANIA, Swiss Re.

[17] indicator that compares claims cost and operating expenses to premium.

denominator of both expense ratio and loss ratio, its decrease causes the ratios to get closer to 1. As previously explained in paragraph 2.1.2, combined ratio is simply the sum of these two ratios.

Motor insurance business presents a very high combined ratio fluctuating around 1. The technical performance of this business follows cyclical periods of profit and losses. Overall, it does not have a crucial direct impact on profitability and Return on Equity[18] index. The real gains from this business are enormous volume of liquidity and establishment of relationships with new customers. Examples of benefits for establishments of new relationships are the opportunities of creating new business and cross-selling of insurance policies of other regulatory classes.

2.2.2. Evolving Customer Needs

Nowadays markets are characterized by high intensity of competition, globalization and digitalization (Lambin, 2007). In this context, customers relationship has become crucial for business success (Brondoni, 2009). The customers have evolved and are increasingly demanding: they need products with additional layers of services (Porter & Heppelmann, 2014) as well as an excellent post-sales customer care (Porter & Heppelmann, 2015). The communication between company and customers not only need to be a two-way communication but also multi-channel: physical and online (Brondoni, 2008). Customers not satisfied will have no trouble in starting a new relationship with a competitor, which understands the real value of establishing solid relationships with its customers (Gordini, 2010).

In the motor insurance business, the MTPL policy bound to a telematics device would allow the company to understand customers behaviour, driving attitude, and life style. This new knowledge would allow insurance companies to make a positive impact on the life of those customers because, on the one hand, it would allow them to reduce the risks they undertake, and, on the other hand, it would allow them to pay lower premiums because they are mitigating their risks or even allow them to experience the insurance in a completely different way.

If is straightforward to think about companies offering customers to underwrite insurance policies for a limited period of time depending on where they are geolocated in that moment because the insurer understood they are leaving for a trip. Therefore, customers will receive a notification on their smartphones proposing them a two-days trip insurance policy. These capabilities require insurance companies, on the one hand, to learn to know their customers and create a customer experience simple but effective, and, on the other hand, to protect from risks that would otherwise have been excessively expensive or would have not been discovered.

2.3. VEHICLE TELEMATICS

Vehicle telematics refers to insurance policies that are bound to a telematics device. The telematics device is connected to the insurance company and it allows to monitor behaviour of the insured vehicle. In Italy, the most adopted devices for vehicle telematics are telematics black-boxes installed in cars, however, even a simple smartphone could be used as telematics device. Generally, a telematics device is composed of:

[18] Measures the profitability of a business in relation to the total equity.

- Internal memory to store the data;
- Battery to ensure functioning in the event of a power failure;
- GPS module to generate GPS coordinates data to track vehicle position;
- Sensors – such as tri-axial accelerometer, gyroscope, etc – to generate additional data;
- GSM/GPRS module to transmit data to the platform and localize the vehicle by cell towers triangulation.

The GSM/GPRS module uses a GSM card to connect the internet and telecommunicate the data; the GSM card is provided by a telecommunication company and has monthly fee and band limit. The quality of information gathered depends on the variety of sensors installed in the device and on the sampling rate of the data. Several different types of sensors guarantee more data sources – e.g. gyroscope, accelerometer, GPS. Higher sampling rate reflects in more data gathered per trip. Obviously, more expensive telematics black-boxes and larger data volume to transmit through internet reflects on higher investment to gather telematics data.

2.3.1. Business Opportunities

Italy is global leader and pioneer in telematics applied to car insurance (Carbone, 2016). According to ANIA (2017) approximately 4.7M out of 30M cars have installed a black-box. Intrinsic characteristics of Italy and its insurance market were breeding ground for adoption and development of vehicle telematics (Dang, 2017).

It is possible to identify at least four levers of value creation related to this technology applied to insurance business: (1) Risk selection, (2) Pricing risk-based, (3) Loss Ratio improvement, and (4) Value-added services and new business models.

Risk selection

Vehicle telematics made its first appearance as a risk selection tool[19]. This was the first application fostering the spread of telematics products in Italy (Carbone, 2015).

In 1999, according to ANIA motor insurance market lost 16% of total premium volume and insurers increased tariffs by 16.7%[20]. Tariffs had increased due to high claim frequency and high average claim cost. By May 2003, the Government declared a Protocol which main objectives were (1) to counter root-causes underpinning tariffs sharp increase and (2) to foster road accidents prevention and reduction of average claim cost on the long term.

It was in this context that vehicle telematics appeared to insurance companies as a risk selection tool, allowing to select risks at the underwriting stage. In fact, telematics products were heavily discounted attracting “good” customers characterized by virtuous and safer behaviour. “Risky” customers were not interested in purchasing this type of product because of the steady monitoring.

The marketing strategy of Sterling Insurance was to offer telematics products in Southern Italy territories where the “average” customer was profiled as “risky”. High risk for insurers reflected in

[19] Source: *Assicurazione Italiana 2003-2004* published by ANIA

[20] In opposition to a general inflation of 1.7%.

high tariffs for customers and this presented a very strong potential of discrimination between "good" and "average" customers. Vehicle telematics became the enabler to grasp this very strong potential of discrimination by guaranteeing discounts to customers purchasing vehicle telematics products. The "good" customers were the niche target, characterized by safer driver behaviour and higher civic responsibility. Telematics products appealed them by guaranteeing considerable price difference compared to traditional insurance products. Discounts strategies were possible because of the significant improvement in technical performance mainly related to claims frequency.

Pricing risk-based

Through vehicle telematics insurance companies became able to monitor risk exposition throughout the coverage period. Data gathered by monitoring the insured vehicle allowed to measure risk of the single customers. It became possible to incorporate this information into parameters to build tariff reflecting an individual pricing of the risk insured.

By the second half of 2000s, telematics products were no more solely related to an up-front flat discount, but they started becoming Usage-based insurance (UBI). The price of UBI products is dependent on distance (Pay As You Drive), on time (Pay When You Drive), on behaviour (Pay How You Drive).

Loss Ratio improvement

In Section 2.1.2 "Key Performance Indicators" it was explained Loss Ratio is a core indicator measuring whether a company is paying more claims than the amount is collecting in premiums. Total claims expenses are obtained by multiplying total claims frequency by average claims cost.

When vehicle telematics first appeared as risk selection tool to improve risk underwriting, it mainly improved claims frequency because "good" customers were characterized by safer driving behaviour and no frauds.

Today, vehicle telematics has the potential of allowing insurance companies to build proactive and fast claims management process. This represent a considerable margin of technical improvement by reducing average claims cost. Improvement of claim handling process would reduce costs and improve Loss Ratio. Additionally, more efficient claim handling would generate value by enhancing customer experience.

Value-added services and new business models

The concept of services is a broad topic that needs to be extended to automotive industry. The new generation of cars is featured with Advanced Driver-Assistance Systems These systems are also referred to as Active Safety Systems because the car itself takes action to prevent or minimize crashes and crash damages. Safer cars imply less premiums for insurers, and, on the other hand, car manufacturers represent a possible threat of new entrants[21] for the opportunities presented by vehicle telematics.

Today, car manufacturers are one of the risks of potential disruption for the car insurance industry. The point is that with advancing of these technologies the role of insurer changes: it cannot be anymore solely undertaking the responsibility of a person against civil liability. New incumbents will

[21] Porter, M.E. (March–April 1979) *How Competitive Forces Shape Strategy*, Harvard Business Review.

enter the car insurance market; they could be car manufacturers, start-up's or giants of other industries. The conclusion is that business model of car insurance is facing several risks potential disruptions. Opportunities offered by vehicle telematics represent one of the possible answers to the need of innovation.

This scenario is very reasonable in the mid-long term. Exactly here stands the vision of Sterling Insurance: to develop vehicle telematics for improving customers relationships, increase loyalty and offering new, high value-added services answering needs of its customers. When it comes to these services, today some of the potential opportunities offered by vehicle telematics can be grouped in three macro categories:

- Services related to the insurance policy, typically delivered through smartphone app. These services can include insurance policies sold "on the go" or tariff adjustment on renewal based on the driving behaviour – Usage Based Insurance like Pay How You Drive, etc.
- Services related to customer's own car. For example, services of car diagnostic.
- Services related to customer's journey. While driving, services can include bad weather alert, speeding alert and even alert when car leaving a pre-defined "safe area". In the event of crash, examples can be services related to immediate assistance and simplified procedure of claim settlement. When parked, locate the car in the case of theft and send alert when the car is moved or damaged.

2.4. STERLING INSURANCE

Sterling Insurance Group is a multinational insurance and banking group leader in Europe and operating in 12 countries involving 32 600 employees. Sterling Insurance is the Italian branch of Sterling Insurance Group and is a middle-sized insurance company.

Accounting for the most important foreign market of the Group, Sterling Insurance has a turnover of €1 500M (2016) and a network of over 1 000 insurance agents distributed on a capillary basis in the territory. By 2016, the company has 820 employees and over 1.6M customers.

Premiums acquired at 31st December 2016 decreased by 9.2% mainly due to shrinkage of Motor vehicle liability business, partially offset by increase of non-motor policies. The turnover breakdown is Property and Casualty business accounting for 73.5% and Life business accounting for 26.5%.

The most important policies for Life business are Whole and term life insurance accounting for 81.1%. As far as concerns Property and Casualty business, Motor vehicle liability accounts for 54.3% and Land vehicle policies account for 7.9%; other most relevant policies are Personal accidents, General liability, Fire and natural losses and Other damage to property.

Considering these figures, the core business of Sterling Insurance is car insurance. However, the management strategy is aiming to diversify the Property and Casualty business by improving the mix between motor and non-motor insurance policies because the company is very exposed to the risk of downturn or potential disruption of the motor insurance market.

2.4.1. Motor Line of Business

By 31st December 2017, Sterling Insurance' portfolio for Motor vehicle liability policies accounts for 1.09M customers and the 68.2%[22] is concentrated in Central-Southern Italy. It becomes clear why the company has been a pioneer in offering telematics products in order to meet the needs[23] of its customers in Southern Italy. Currently, 350 000 policies are telematics products and the goal are to reach 600 000 in the next two years.

Sterling Insurance' offer to market is constantly evolving. The main car insurance products are *Guidamica* and *Autocontrollo*. *Guidamica* is the traditional motor vehicle liability policy. *Autocontrollo* is the telematics product requiring customers to install a telematics black-box inside their vehicle. On February 2018, *Autocontrollo* grants an immediate 18% discount on the tariff and up to 32% discount in the following years, depending on car usage[24].

Additional services offered for telematics products rely on the smartphone app *MyAngel* allowing customers to localize the car in the event of theft, to receive crash assistance, to access towing service 24/7, and to visualize summary reports to monitor their usage behaviour. This app is also a very important asset to answer the need of evolving customers as previously explained in Section 2.2.2.

The shrinkage of motor premiums compared to previous year (2015) was caused by a decrease in the average premium of Land vehicle liability equal to -6.17%, recorded with the same magnitude by the entire Italian market[25]. The decrease in the average premium was caused by stronger competition eroding the market profitability as well as by spreading of telematics policies sold at a steep discount.

Sterling Insurance' vision about motor insurance business revolves around the world of services. The management believes in Telematics becoming a tool to improve knowledge of its customers' needs and establish solid relationship with them. This scenario would present the fundamentals of an entire new world of opportunities and possibilities.

2.4.2. Corporate Strategy

The strategy "Telematics and Internet of Things" of Sterling Insurance aims to embrace the change represented by: evolution of the consumer, new varieties and massive amount of data generated in real time, and high intensity of competition exposing the business to the constant risk of potential disruption of other industries or new technological incumbents.

The main objective of Sterling Insurance is to increase the knowledge about its customers profile and behaviour in order to establish more solid and longstanding relationship with them. This aspect is crucial for two reasons: first, nowadays customers' need additional layers of services integrating the core products and demand for multi-channel communication; second, vehicle telematics requires customers to trust the company in order to allow sharing their personal data.

[22] Source: Sterling Insurance Planning Office

[23] As explained in section 2.1, telematics products allowed good customers in Southern Italy to pay a fair price for their risk

[24] *Autocontrollo* product has Pay As Your Drive features

[25] Source: ANIA

The strategic objective is to change the relationship with customers: it will be no more solely an interaction in the event of claim, but it will become a relationship for risk prevention, suggestion for improving specific behaviour and even offer services to answer customers' needs.

The increase of knowledge to better satisfy customers' is bound to the development of analytics capabilities required to analyse Big Data and deploy in operations the discovered insights. Obviously, the trust in sharing personal data remains the *conditio sine qua non* for developing these opportunities.

Developing analytics capabilities for Big Data is a complex goal to achieve because it requires distributed storage architecture capable of storing around 1 000 TB per year, a computational engine to work on a distributed storage environment, and the employees capable of using these tools to extract knowledge from Big Data useful for the business. Concerning the implementation of these technologies, Sterling Insurance invested on purchasing two interconnected cloud platforms – which are G-Evolution assets. One platform ingests raw data in near-real time from telematics black-boxes and stores them in data lakes. The other analyses such massive amount of data using a distributed storage and computing framework. Unfortunately, these platforms were not available when this project was carried out.

With the first quarter of 2018, the company successfully achieved its objective of increasing the portfolio of telematics products to 350 000 and deploying the Big Data storage and analytics platforms.

2.4.3. The Internship

The internship is part of the "Internet of Things and Telematics" strategy introduced in previous section. It is carried out in the Actuarial Department at the company Headquarters. The Actuarial Department is the most important department for an insurance company because it has the technical know-how to drive the company and to foster innovation.

The strategic objective of the internship consists of performing research and development around trips raw telematics data in order to build and share technical know-how within the whole company. Trips raw telematics data refers to data at the granular level of the GPS point generated by telematics black-boxes during trips; the distinction is important because there are also raw telematics data generated during crashes.

The technical know-how developed is related to all those analytics approaches and data science methodologies required to build a data science solution according to CRISP-DM methodology. The project articulates from the understanding of the business problem and definition of data mining tasks to the extraction of a new level of knowledge about customers behaviour.

The internship is important because deals for the first time with telematics data at the granular level of geospatial coordinate set. The results discovered represent the starting point for the deployment for concrete and operational uses of Telematics data.

This chapter has tried to summarize the exciting context surrounding the work and the domain knowledge developed throughout the internship. The next chapter describes in detail the methodology and the technical development of the project.

3. DATA AND METHODS

In this research, we adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) method (Shearer, 2000). The most important aspects of this methodology are the centrality of business understanding and the inter-dependent and cyclical nature of the its six phases. The latter reflects in the results of an entire iteration of the process to become the insights for improving specific stages or task and iterate the whole process again.

In this chapter are extensively described tasks of all phases of CRISP-DM process except for the deployment phase which is treated in the next chapter.

Data understanding and business understanding phases have been iterated copious times allowing to narrow the focus of the business problem and understand in-depth the data – at the granular levels of coordinates, segments and entire trips.

The data preparation phase of CRISP-DM has been divided into two phases: data preparation and data pre-processing. Data preparation phase encompasses all tasks required to obtain the structure of the final dataset starting from the raw data. Data pre-processing phase includes all tasks applied to obtain the final dataset to feed the clustering algorithm. The raw data are at the granular level of the GPS point while the final dataset is aggregated at the granular level of the entire trip.

In the data preparation phase, trips raw telematics data at the granular level of GPS coordinates is cleaned, enriched with external data sources and aggregated to the granular level of segments. Then, for each segment are computed features to describe its characteristics – for example average speed and road type. The next step consists of aggregating this secondary data at the granular level of segments to obtain the structure of the final dataset at the granular level of the entire trip.

Once obtained the structure of the final dataset – aggregated to the granular level of entire trips – an extensive data pre-processing phase is accomplished. It required great efforts in terms of time including features engineering, noise filtering, coherence checking and outliers' treatment. Then, the final dataset is obtained and feed to the modelling phase. Variable selection and dimensionality reduction tasks were cyclically improved with the insights discovered evaluating clustering results.

The modelling phase consists in learning a clustering model over trips characteristics using the k-means algorithm on the final Analytical Base Table – outputted from the data pre-processing phase. The modelling phase is strictly interdependent with data pre-processing tasks – such as outliers' treatment, features engineering, variables selection and dimensionality reduction – and the evaluation phase – which consists in thoroughly evaluating the model before deploying it.

The insights gained during each iteration of the evaluation phase are incorporated in Section 3.5 where is described the clustering model. The output of modelling phase is extensively studied adopting a clusters profiling framework inspired on the functionalities offered by SAS Enterprise Miner.

In the next chapter are discussed the results of the final clustering model and its results are deployed to profile each of the 937 customers.

3.1. BUSINESS UNDERSTANDING

The first phase of the CRISP-DM is business understanding. It focuses on understanding the project objectives from business perspectives and converts this knowledge into a clear defined data mining problem. This allow to develop a preliminary plan to achieve the goals. Considering the nature of the CRISP-DM methodology, this phase is cyclically performed several times and each iteration improves the understanding and better focuses the business goal of the project.

The business question to answer is: how Sterling Insurance can analyse raw telematics data and whether it can discover useful knowledge. Considering the Research and Development context of the current project, understanding of the business and problem definition have evolved as the project progressed and it discovered specific aspects to focus on. The problem to address has evolved and narrowed its scope as understanding and know-how about raw telematics data increased. Initially, the objective was broad aiming to discover some hidden structures from trips telematics data using unsupervised machine learning techniques. Obviously, the patterns need to be described from the perspective of insurance business, hence, they are intrinsically related to risk features of trips.

By the end, the primary business goal of this project has converged toward clustering all trips in the dataset based on their summary characteristics and regardless of drivers historical and personal information. As the secondary business goal, the knowledge derived from the overall portfolio is applied to each customer to discover its most recurrent patterns and define the unique profile of each driver.

The suitable measures of success for the solution are rather qualitative than quantitative. This is due to the Research and Development nature of this project. The quality of the clustering is not determined using statistical indices which are not relevant for the business, instead expert judgement and domain knowledge are used to determine the relevance of the patterns discovered.

3.2. UNDERSTANDING TELEMATICS DATA

Trips raw telematics data used for this project were generated by sensors and GPS module embedded in telematics black-box devices. The data have been transmitted by the GSM/GPRS module through network connection toward the cloud platform of S-Evolution[26]. The sampling rate of data collected is influenced by the bandwidth cap of the GSM card connecting to internet. The data used for the present study has a sampling rate of one GPS point approximately every 2000 meters and one GPS point every time a *behavioural event* is triggered. The fundamental difference between *position point* and *behaviour event* is explained in Section 3.2.1.

When referring to telematics data, it is important to distinguish between aggregated data and raw data. Telematics aggregated data are at the granular level of the entire day. They are pre-processed and enriched by the telematics black-box manufacturer[27] in order to present daily summary statistics to Sterling Insurance. The reader should think of this type of telematics data as total kilometres travelled throughout the day broken-down into time slots – e.g. morning or evening – and road types

[26] Remembering Section 2.1, S-Evolution is a company founded by Sterling Insurance which corporate purposes are to offer high added value services based on vehicle telematics.

[27] Octotelematics is the global leader in providing vehicle telematics and data analytics solutions to insurance industry.

– e.g. highway or secondary roads. These daily reports enrich Sterling Insurance historical knowledge about customers' profiles, feeding the data warehouse for weekly, monthly and yearly summaries. The *telematics score* improving pricing is computed starting from this information.

Instead, trips raw telematics data are something new. The level of granularity is single geospatial coordinate set composing trips trajectories. Sterling Insurance refers to this data as "raw" because, when compared to aggregated data, they have not been subjected to data enrichment, pre-processing nor aggregation stages. Handling raw telematics data requires Sterling Insurance to develop capabilities, implement technologies and build ETL processes required by this new type of data. As explained in Section 2.3 "Vehicle Telematics", the value of this data is potentially enormous unlocking new knowledge, opportunities and business models.

This project contributes to the design of processes and development of capabilities to deal with raw telematics data. It creates internal know-how about managing this data and creates an entire data science pipeline going from the flat file to the business application of resulting knowledge discovered.

3.2.1. Data Source

The current project is a Proof of Concept and the data sources are several flat files supplied in the form of batch extractions from Octotelematics data warehouse. The only difference between this project and the operational scenario is the data source which instead will be a data lake on the cloud. However, considering data structure remains the same, the methodology to analyse raw telematics data does not change. This allows to design the processes Sterling Insurance need when the Big Data platforms will be deployed.

For completeness of the study, it is interesting to describe the raw telematics data source in the operational scenario using Cloud Big Data Platforms. First, raw telematics data are generated by telematics black-boxes. Next, they are sent to OCTO Telematics Platform[28]. OCTO Telematics Platform do not enrich the data with only exception for customers contract identification (voucher id). Finally, OCTO Telematics Platform stream in semi-real time the data to Sterling Insurance' storage platform. The data source will be distinguished in three different flows feeding three data lakes: *Positions & Trips*, *Behavioural Events* and *Crash*.

Position & Trips flow data are referred to as *position points*. They are generated when the engine is turned on or turned off and approximately every two kilometres. Each *position point* consists of one GPS position, timestamp[29], instantaneous speed, distance travelled since previous *position point*, and few other. It is critical to bear in mind telematics black-boxes take into account exclusively *position points* when computing the distance travelled since the previous *position point*.

The Behavioural Events flow data are referred to as *behaviour events*. They are generated in the case acceleration thresholds are activated triggering a *behavioural event*. Two important differences between *behaviour points* and *position points* are: (1) *behaviour points* are integrated with summary acceleration information gathered from 3-axis accelerometer sensor; (2) *behaviour points* are not considered for computing distance travelled since the previous position.

[28] OCTO Telematic Platform is the cloud platform of the telematics black-box manufacturer.
[29] A timestamp is the time at which an event is recorded by a computer.

The Crash flow is generated when OCTO Telematic Platform detects a crash event. It consists of one position record enriched with a time series of acceleration data gathered from 3-axis accelerometer sensor.

The current study neglects *Crash* flow and uses only *Position & Trips* and *Behavioural Events* flows. Unfortunately, the batch extraction of Octotelematics did not include the summary acceleration data of *Behavioural Events* flow. This is a considerable limitation for the project because it is not possible to use the summary acceleration data for interpreting the real nature – i.e. hidden behind each *behaviour event.*

Figure 3.1 represents a possible scenario where a monitored car makes trips and remain parked between a trip and the next one. This figure is very important to understand because the concepts it visualized, are referred throughout the report.

Each trip is composed of a sequence of GPS coordinates, timestamp and metadata[30]. Each trip starts with the Ignition On event and ends with the Ignition Off event – respectively green and red diamond shapes in Figure 3.1.

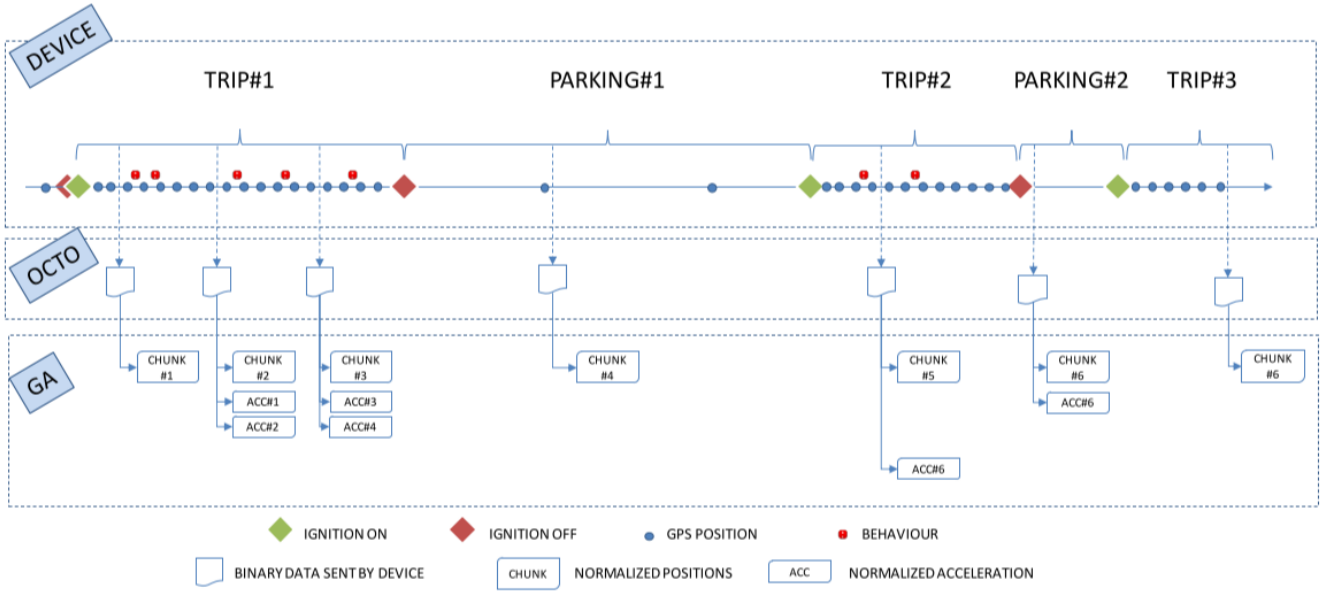


Figure 3.1 – Example of creation and transmission of trips raw telematics data

Throughout the trip, telematics black-boxes generate *Position & Trips* data – represented by the blue circles in Figure 3.1 – approximately every two kilometres. The *Behavioural Events* data – represented by the red circles in Figure 3.1 – are generated when acceleration thresholds are triggered.

The parking time is the time elapsed between the timestamp of the Ignition Off event in the previous trip and the timestamp of the Ignition On event in the next trip. Even though acceleration data have been generated, the data source is not available for this project.

[30] Metadata providing information about coordinates and timestamp are speed, heading, GPS quality and distance travelled since the previous location.

3.2.2. Data Structure

The data are collected through a batch extraction of all trips made from June 2017 to November 2017 by 937 customers. The batch extraction consists of several flat files[31] containing a combination of *Position & Trips* flow and the position data of *Behavioural Events* flow – accelerometer data was not included in the batch extraction, hence, not available for the study. The lack of accelerometer data of *Behavioural Events* flow is a great loss of information for the study. Without accelerometer data it is not possible to understand causes and to interpret the behavioural event: it is known the behavioural event was caused by some acceleration thresholds triggered but the nature of the event – harsh acceleration, harsh break, etc. – remains unknown.

In Figure 3.2 is presented an example of the flat files as they were received.

```
2;13585989;358639059380330;28/07/2017 12:28:43;44.533550;10.863609;0;160;3;0;8045762;0
2;13585989;358639059380330;28/07/2017 12:33:02;44.532467;10.875152;30;38;3;2051;259;1
2;13585989;358639059380330;28/07/2017 12:35:37;44.546761;10.885636;36;22;3;2002;155;1
2;13585989;358639059380330;28/07/2017 12:38:24;44.551842;10.902075;41;8;3;2032;167;1
2;13585989;358639059380330;28/07/2017 12:40:14;44.570116;10.905954;32;6;3;2058;110;1
2;13585989;358639059380330;28/07/2017 12:43:09;44.587214;10.913594;39;18;3;2006;175;1
2;13585989;358639059380330;28/07/2017 12:45:24;44.600603;10.930426;32;42;3;2025;135;1
2;13585989;358639059380330;28/07/2017 12:46:45;44.610366;10.936056;6;48;3;0;0;1
2;13585989;358639059380330;28/07/2017 12:46:48;44.610458;10.936262;20;54;3;0;0;1
2;13585989;358639059380330;28/07/2017 12:47:16;44.613405;10.936774;17;38;3;0;0;1
2;13585989;358639059380330;28/07/2017 12:47:21;44.613553;10.937216;19;44;3;0;0;1
2;13585989;358639059380330;28/07/2017 12:47:54;44.617053;10.940978;48;44;3;2092;150;1
```

Figure 3.2 – Example of raw telematics data flat file

In Table 3.1 is presented the structure to apply the flat files for structuring them into Analytical Base Tables. In this structure, each record contains data related to a single GPS location.

Field	Datatype	Description
ID	Unary	Packet identifier equal to 2 for trips data
device_id	ID	International Mobile Equipment Identity (IMEI)
voucher_id	ID	Customer contract ID
trip_key	ID	Trip identification generated when an Ignition ON is detected
timestamp	String	Timestamp in seconds and GMT time zone
latitude	Numeric	Latitude coordinate
longitude	Numeric	Longitude coordinate
speed	Numeric	Speed in mi/h averaged on one second timeframe
heading	Numeric	Direction of GPS motion in degrees compared to the North
gps_quality	Categorical	GPS signal quality (0 = no signal, 1 = navigation in 2D, 2 = marginal signal, 3 = navigation in 3D)
distance_elapsed	Numeric	Meters from the previous *position point*
time_elapsed	Numeric	Seconds elapsed since previous *position point*
session	Categorical	Engine status (0 = ON, 1 = Movement, 2 = OFF)

Table 3.1 – Structure to apply to the raw telematics data source

[31] In comma-separated values format.

3.2.3. Raw Telematics Data Exploration

In this section is accomplished the preliminary exploration of trips raw telematics data – at the granular level of single geospatial coordinate set. This allows to understand the meaning of each variable and identify potential issues. Descriptive summary statistics for each variable are summarized in Appendix 2.

Once the structure of Table 3.1 is applied to the flat files of Figure 3.2, the input data for the data mining process is obtained. In this Analytical Base Table there are 13,821,131 records and 11 variables[32]. Each record contains information about a single geospatial coordinate set of the trip. Geospatial coordinates can belong to Position & Trips flow or to Behavioural Event flow.

In Table 3.2 it is shown as example one trip from the input Analytical Base Table. A value of variable "session" equal to 0 indicates the Engine On event, hence, the trip starts. Values equal to 1 for this variable indicates the car is travelling and when the value is equal to 2 it indicates the Engine Off event, hence, the car is turned off and the trip ends. Variable "time_elapsed" presents high value in the very first record because it is Engine On event meaning the car had remained parked 93 days.

trip_key	voucher_id	device_id	timestamp	latitude	longitude	speed	heading	gps_ quality	distance_ elapsed	time_ elapsed	session
1	15358898	358639	7/28/2017 14:28	44.******	10.******	0	160	3	0	8045762	0
1	15358898	358639	7/28/2017 14:33	44.******	10.******	48.28	38	3	**2051**	259	1
1	15358898	358639	7/28/2017 14:35	44.******	10.******	57.94	22	3	**2002**	155	1
1	15358898	358639	7/28/2017 14:38	44.******	10.******	65.98	8	3	**2032**	167	1
1	15358898	358639	7/28/2017 14:40	44.******	10.******	51.50	6	3	**2058**	110	1
1	15358898	358639	7/28/2017 14:43	44.******	10.******	62.76	18	3	**2006**	175	1
1	15358898	358639	7/28/2017 14:45	44.******	10.******	51.50	42	3	**2025**	135	1
1	15358898	358639	7/28/2017 14:46	44.******	10.******	9.66	48	3	0	0	1
1	15358898	358639	7/28/2017 14:46	44.******	10.******	32.19	54	3	0	0	1
1	15358898	358639	7/28/2017 14:47	44.******	10.******	27.36	38	3	0	0	1
1	15358898	358639	7/28/2017 14:47	44.******	10.******	30.58	44	3	0	0	1
1	15358898	358639	7/28/2017 14:47	44.******	10.******	77.25	44	3	**2092**	150	1
1	15358898	358639	7/28/2017 14:48	44.******	10.******	41.84	54	3	0	0	1
1	15358898	358639	7/28/2017 14:49	44.******	10.******	32.19	50	3	0	0	1
1	15358898	358639	7/28/2017 14:49	44.******	10.******	82.08	44	3	**2017**	115	1
1	15358898	358639	7/28/2017 14:49	44.******	10.******	57.94	34	3	0	0	1
1	15358898	358639	7/28/2017 14:50	44.******	10.******	25.75	120	3	0	0	1
1	15358898	358639	7/28/2017 14:56	44.******	10.******	0	0	3	**567**	396	2

Table 3.2 – Example of raw telematics data of one trip

Variables "distance_elapsed" and "time_elapsed" are computed exclusively taking into account *position points* and neglecting *behaviour events*. Values equal to 0 simultaneously in variables "distance_elapsed" and "time_elapsed" indicate the record is related to a *behaviour event* triggered by some acceleration thresholds.

Variables "latitude" and "longitude" present five decimals and give GPS positions up to 1.1 meters.

Once the data at this granular level are validated, they will be enriched with context and, then, aggregated at the granular level of the single trip. The objective is to obtain an Analytical Base Table in which each record represents a single trip and each column represent a feature describing the characteristics of the trip.

[32] As it will be explained later, variable "ID" was dropped because it is a unary variable.

ID

This field is a string used by Octotelematics to identify package of information transmitted. It guarantees that different data flows have different ID.

In the case of Position & Trips flow, variable ID is unary valorised 2. Hence, it has no utility for this analysis and it was dropped.

device_id

This variable contains the IMEI number to univocally identify the telematics black-box. The International Mobile Equipment Identity is a unique number used by a GSM[33] network to identify the device.

The dataset has 939 unique Device_id.

voucher_id

This variable contains the unique ID of the customer contract. It is important to state this primary key does not provide a direct lookup with internal databases of Sterling Insurance containing personal information about customers. Therefore, it is not possible to enrich the data source with additional information such as customer's age, vehicle's characteristics, etc.

In Figure 3.3 it is visualized the total number of trips performed by each customer arranged in descending order of frequency. It can be observed the maximum number of trips performed by a single customer is 3019, the first half of the portfolio performed more than 1000 trips and the last quarter of the portfolio performed less than 500 trips.

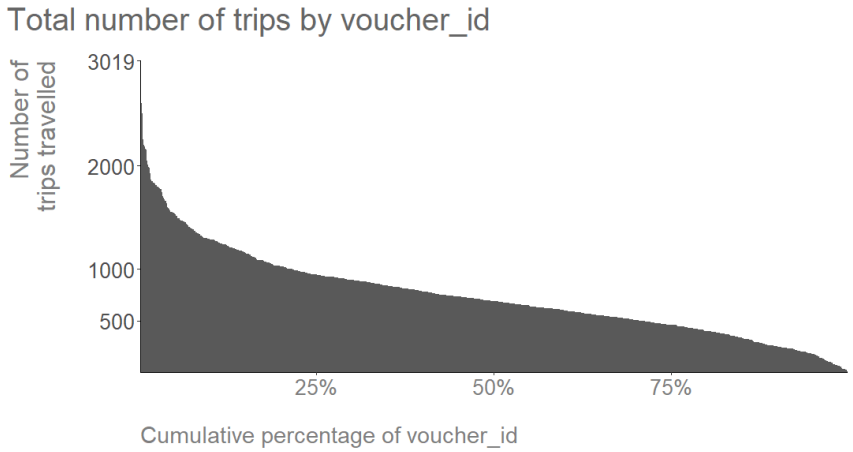


Figure 3.3 –Total number of trips travelled by each voucher

The dataset has 937 unique "voucher_id", hence, 937 unique vehicles. The fact there are two more device_id than voucher_id can be explained by the fact that telematics black-boxes installed in the cars of two customers were replaced.

Variable "voucher_id" will be used to study the recurrent patterns of each customer and define its specific profile.

[33] Global System for Mobile Communications.

trip_key

Variable "trip_key" is the primary key to univocally identify each trip in the dataset. As explained in Figure 3.1, a new trip starts when an Ignition On event is detected. The dataset contains 696 953 trips, hence, variable "trip_key"

In Figure 3.4, it can be observed the number of trips for each month. November has the least number of trips because the timeframe of the dataset ranges from 1st July 2017 to 21st November 2017. The timeframe of individual customers can differ from the timeframe of the dataset.

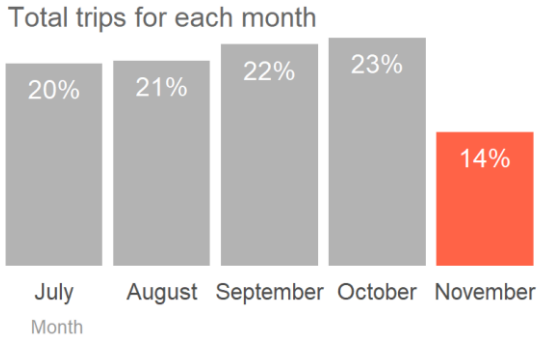


Figure 3.4 – Relative frequency of trips by month

It is interesting to observe that 41.13% of the trips were in the summer period. This indicates presence of seasonality caused by summer trips or by behaviours that differ from the ordinary customers routine. These aspects require investigation.

Variable "trip_key" will be used to aggregate the initial dataset of raw telematics data until collapsing into a dataset composed of one single record describing each trip using features to effectively summarize its characteristics.

timestamp

Variable "timestamp" is a string with granularity to seconds. It contains information about the time when the record was generated by the telematics black-box. The time zone is Greenwich Mean Time[34], therefore, it needs to be adjusted to the correct time zone for Italy[35] taking into account the Daylight Saving Time.

Variable "timestamp" might be expected to be unique for each customer. On the opposite, it was discovered all duplicated timestamp values for each customer there sum up to 83 691 records. This phenomenon affects 10.14% of the trips and requires further investigation in order to understand its causes and how to handle it.

To deal with this variable in *R* language, the string is converted into a date/time class. There are two possibilities: (1) POSIXlt stores a list of year, month, day, hour, minute, seconds and other time information; (2) POSIXct stores the signed number of seconds since the beginning of 1970 in GMT time zone. The latter requires less memory and was the format chosen to store this variable.

[34] Not to be confused with British Summer Time which is the time zone of United Kingdom.
[35] Central European Time.

From this variable can be obtained dimensions related to the time length of trips, as well as different granular level of temporal information such as month, week of the year, day of the week and time slot.

latitude and longitude

Variables “latitude” and “longitude” contain geographical information about the GPS coordinate when the record was generated by the telematics black-box.

In Figure 3.5 are visualized latitude and longitude coordinates of the entire data set. Most of the trips were performed in Italy but there are several trips abroad extending from Morocco to Romania. The geolocation of 11 records is in latitude 0.00 and longitude 0.00 and this error needs to be investigated.

Figure 3.5 – Visualization GPS coordinates on map

According to the framework for data mining applied to spatial trajectories, GPS coordinates should be pre-processed using Noise Filtering and Map-Matching techniques. The first are necessary because spatial trajectories can be inaccurate due to poor positioning signal. The latter are intended to convert a sequence of raw GPS coordinates to a sequence of road segments. However, the use of spatial trajectories is out of the scope for the project because this study exclusively focuses on the summary characteristics describing trips.

Combining open data and GIS methodologies available in several *R* packages developed by E. Pebesma, it is possible to obtain useful geographic data and integrate trips with relevant context. Great efforts are spent in this task and are discussed in the Data Enrichment Section 3.3.2.

speed

This variable represents the average speed in mi/h computed on the timeframe of one second gathered when the record was generated by the telematics black-box. Before proceeding the variable is converted to km/h.

To better visualize speed distribution in Figure 3.6, all values above 150 km/h up to a maximum of 288 km/h have been filtered (0.14% of records). There are two important insights: first, there is a concentration of 19.31% records featured with a speed equal to 0 km/h which seems suspicious. It might be partially explained by the fact every trip starts and ends with a value of variable "speed" equal to 0 km/h; anyway, it requires investigation.

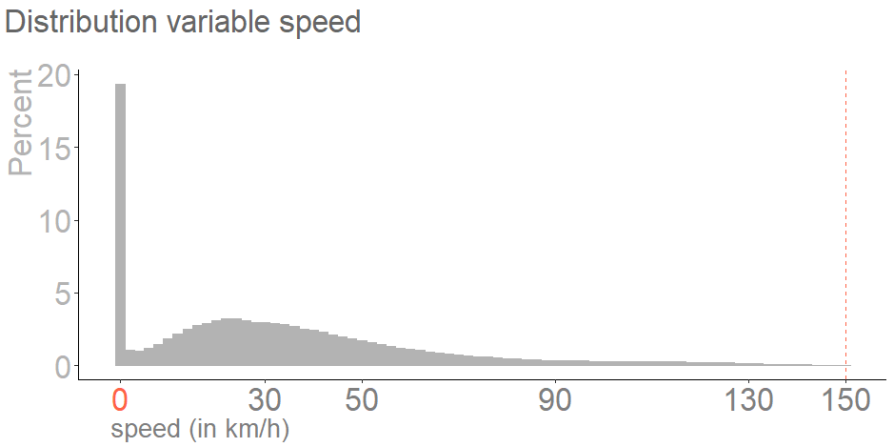


Figure 3.6 – Distribution variable *speed*

Second, the 0.14% of records present a speed above 150 km/h with a maximum of 288 km/h. It should be investigated whether they are measurement errors and, eventually, the conditions causing such errors.

heading

Variable "heading" measures the direction of the car in degrees compared to North. It is a circular variable meaning the minimum is also the maximum value; 0° and 360° are in the same position. Therefore, it should present range between 0° and 359° or between 1° and 360°. Additionally, this variable is expected to be evenly distributed.

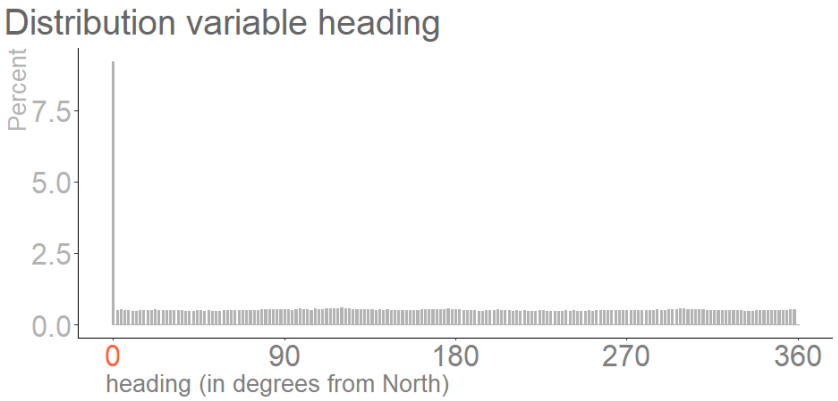


Figure 3.7 – Distribution variable *heading*

Observing Figure 3.7, it is possible to gain two insights. First, the variable ranges from 0° to 360°. Second, there is a peak of 9.21% of records featured with a value of 0°. Combining these two insights it starts becoming clear that 0° does not represent degrees from North but it is a Missing Not At Random. In fact, Octotelematics – the provider of the data – confirmed that it this is a value generated by the telematics black-box when it was not possible to compute the direction of the car.

At this point, it is likely that this fact can contribute to explain the abnormal peak of variable “speed” for the value of 0 km/h.

From this variable can be obtained information about turns of the car, however, the low sampling rate of the data collected most likely causes this information to not be very useful.

gps_quality

This variable measures the GPS quality associated to the record. This is a very important information because it measures the reliability of the data collected. For example, records with poor GPS quality might contain errors in coordinates, variables “speed” or “heading”. The implications of poor GPS quality are studied in detail and handled in Section 3.3.1 “Data Quality Assessment”.

This variable can be treated as ordinal or encoded as numeric from 0 to 4. There are four levels ranging from “No Signal” to “3D Fix” – representing the best possible GPS quality.

In Figure 3.8 it can be observed the lowest level is almost not present in this dataset (only 377 records). A “Marginal Signal” – third highest level – is present in 11.93% of records; Octotelematics stated this is an acceptable proportion. The 3D Fix quality is present in 87.74% of records.

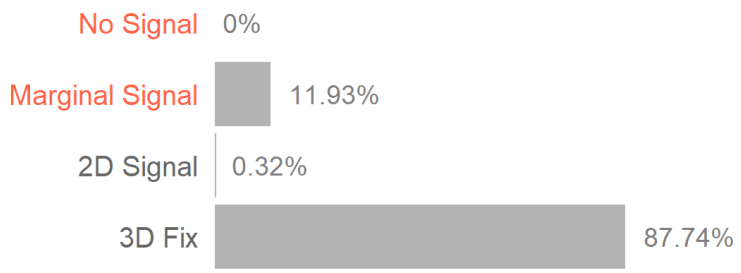


Figure 3.8 –Distribution variable *gps_quality*

From this variable can be extracted information about the overall reliability of trips as well as its variability throughout the trip.

distance_elapsed

Variable “distance_elapsed” represents the number of meters travelled since previous position. It plays a crucial role in this project for several reasons. It makes possible to compute the overall road distance travelled for each trip and the road distance travelled between *position points*. This information underpins the creation of variables based on distance travelled such as average speed, total road distance travelled on broken down for each road type, etc.

In Section 3.2.1 “Data Source”, it was explained the fundamental concept according to which telematics black-boxes exclusively consider *position points* for computing distance travelled between two GPS coordinates. In fact, GPS coordinates related to *behavioural events* are not taken into account to compute “distance_elapsed” representing the meters travelled since previous position.

Therefore, variable "distance_elapsed" helps to distinguish the nature of records, because in the case of *behavioural events* the variable is always equal to 0 meters[36]. In the case of *position points*, there are *position points* featured with variable "distance_elapsed" equal to 0 meters – for example, the first record of each trip corresponding to the Engine On event.

Remembering the discussion on Figure 3.1, it was theoretically expected this variable to not significantly differ from a value of 2 000 meters. The reason is because the sampling rate of *position points* is approximately every 2 000 meters. In reality, we often observe a different behaviour.

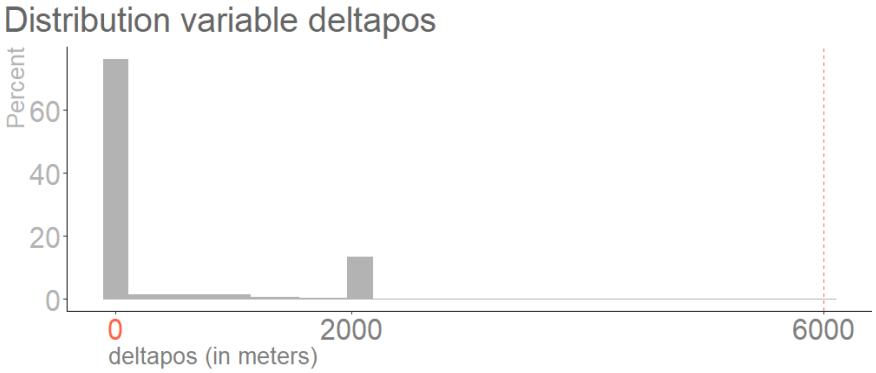


Figure 3.9 – Distribution variable *distance_elapsed*

To allow a better visualization in Figure 3.9, values above 6 000 meters of "distance_elapsed" have been considered as equal to 6 000 meters. They amount to 0.13% of the distribution and range up to a maximum of 65 535 meters; clearly to be investigated and checked for consistency.

The most important insight of Figure 3.9 is the peak of variable "distance_elapsed" in 0 meters. It accounts for 79.79% of records. Remembering that variable "distance_elapsed" is equal to zero in only two cases: Engine On event and *behaviour event*; and that there are 696 786 Engine On events in the dataset – one for each trip. It is simple math to discover records which nature is related to *behavioural events* accounts for 74.72% of the entire dataset.

At this point it becomes even more clear lacking accelerometer data to explain these records is an incredible loss of knowledge for the current study.

time_elapsed

Variable "time_elapsed" represents the number of seconds elapsed since previous position. This variable is only computed by telematics black-boxes considering *position points*, therefore, *behavioural events* are featured with both "distance_elapsed" and "time_elapsed" equal to zero.

There two main differences when comparing "time_elapsed" and "distance_elapsed": 1) regarding the Engine On event – which is the first record for each trip – "time_elapsed" represents the amount of time the car was parked since the Engine Off event of previous trip; 2) variable "time_elapsed" does not have any expected theoretical value as in the case of "distance_elapsed" – which was expected to concentrate around 2 000 meters according to *position points* sampling rate.

[36] According to the insights discovered for variables "heading" and "speed", the value 0 is a Missing Not At Random which is generated in predetermined situations.

Variable “time_elapsed” ranges from 0 up to 99999999 presenting a Skewness value of 743. Therefore, it is not possible to effectively visualize this information since the distribution is right skewed very long right-hand tailed. This behaviour is observed because in reality variable “time_elapsed” contains information related to two different phenomena: 1) inactive time since the Engine Off event terminating previous trip and 2) time elapsed since previous *position point*.

It is appropriate during the data preparation phase to remove the parking time information from variable “time_elapsed” and assign it to a new variable named “inactive_time”.

session

This variable allows to distinguish start and end records of trips – respectively “Engine On” and “Engine Off” – as well as the records belonging to the route – “In Movement”.

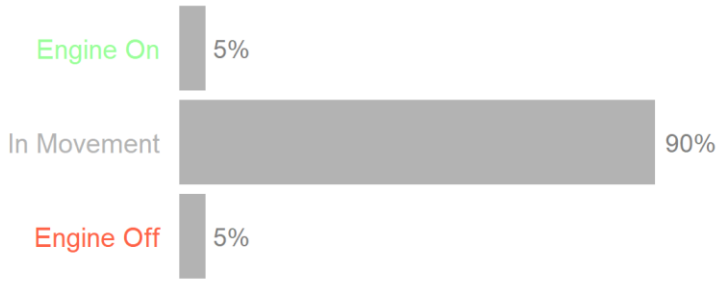


Figure 3.10 – Distribution variable *session*

Observing Figure 3.10, it is chosen the green colour to represent “Engine On” because communicates the idea of start – the start of the trip. It is chosen the red colour to represent “Engine Off” because communicates the idea of end – the end of the trip. The number of Engine On events is equal to the number of Engine Off events which are equal to the total number of trips.

There are no issues to investigate about this variable.

This section is completed and it has presented the summary of cyclical and iterative repetition of Data Understanding phase of CRISP-DM. Next, the Data Preparation phase.

3.3. DATA PREPARATION

In this phase are discussed design and implementation of all processes required to obtain the final dataset feeding the clustering algorithm starting from the raw data. The knowledge database discovery through unsupervised learning is performed over summary characteristics of trips, therefore, the final dataset must be structured as one single record describing each trip.

First, the initial raw dataset at the granular level of geospatial coordinate points is analysed to discover insights and to assess data quality issues. Then, it is enriched with geographical context by combining GIS methodologies and open data. Next, it is created an intermediary dataset at the granular level of segments. The *segment* is a portion of trip between two *position points* for which the road distance travelled – measured by “distance_elapsed” – is available. B*ehaviour events* records are aggregated, and a new variable is created counting the frequency of *behaviour event* in each *segment*.

This approach allows to reliably compute features – such as average speed, distance travelled on motorways, etc. – at a level granularity finer than the entire trip. The decision to not compute this information for each point is driven by the lack of "distance_elapsed" information for *behaviour event* records – which accounts to 74.72% of the records.

Finally, the Analytical Base Table is created by aggregating the information of each trip into one single record. The trips are described creating summary features from the initial raw dataset at the granular level of the GPS point and from the intermediary dataset at the granular level of the *segment*.

To better organize the report, noise filtering, outliers' treatment and feature engineering stages are described Section 3.4 "Data Pre-Processing" because they required to obtain the final Analytical Base Table starting from the output of this Data Preparation phase.

3.3.1. Data Quality Assessment

During the preliminary exploration in Section 3.2 "Understanding Telematics Data", several issues were identified at the granular level of GPS coordinate points. The issues identified are related to variables: "gps_quality", "latitude" and "longitude", "timestamp", "speed", "heading", and "distance_elapsed". These issues might have complex implication and need to be carefully studied and addressed, using the most appropriate approach in the context of this project. Note that some issues will require to be solved at the granular level of trips.

Table 3.3 recaps the result of this data quality assessment in terms of records at the granular level of the GPS point dropped. Not all issues can be solved here, causing the need of dropping entire trips once the raw data will be structured into the final table at the granular level of trips.

Data Quality Issue	Motivation	Affected records
Behavioural events with "gps_quality" level equal to zero	All variables are valorised equal to zero.	181
"gps_quality" level equal to one, "session" in movement and "distance_elapsed" equal to zero	Unreliable behaviour events	1 039 512
Completely duplicated records	This are transmission errors causing to copy a record.	44 842
Duplicated "timestamp", "distance_elapsed" equal zero and "gps_quality" level equal to 1	These records are not reliable. It is not possible to define them as behaviour events with reasonable certainty.	23 330
Missing coordinates for records associated to Engine On event	It is not possible to reliably impute the coordinates	11
Total Removed Records		**1 081 036**[37] (7.83%)

Table 3.3 – Rules defined to remove dangerous issues and results of the cleansing

[37] Obviously, the number of removed records is less than the sum of the affected records because some are affected by multiple quality issues at once.

gps_quality

The first issue to investigate is related to variable "gps_quality": it is important to understand the impact of poor GPS quality on the reliability of information contained in that record. At first glance, it is clear that no major issues are detected for records featured with "gps_quality" higher than level 1 – Marginal Signal.

Investigating the 377 records with "gps_quality" equal to level 0, it is discovered mean and variance of "speed", "heading" and "distance_elapsed" are equal to zero. Among them, 196 records are related to Engine Off event and 181 records are related to the movement session of the car. For obvious reasons the record related Engine Off events cannot be removed. Instead, decision is to filter the 181 records related to movement session because they are not *position points* and the information they contain are very unreliable.

Investigating the 11.93% of records with variable "gps_quality" equal to level 1, it is observed the only numeric variable with mean and variance equal to zero is "speed". Concerning variable "distance_elapsed", it is valorised to zero for 97.00% of these records and, in this subset, 63.07% are related to a motion status of the vehicle[38].

It would be dangerous to indiscriminately drop all records featured with level of "gps_quality" equal to one because they might present delicate issues. In fact, 3.00% of those records are related to *position points* which are the fundamental components of *segments* and trips – representing the records generated every ≈2 000 meters. Additionally, it is difficult to distinguish between records associated to *behavioural events* and ambiguous records with "distance_elapsed" equal to zero.

The decision is to proceed assessing the data quality, bearing in mind the issues related to these records. It is critical to reduce the risk of removing *behavioural* events records because erroneously considered containing unreliable information.

Missing latitude and longitude

There are 11 records missing latitude and longitude coordinates. They are related to Engine On events of 11 different trips made by 11 different customers. The immediate approach is to impute the missing using the GPS coordinates where the previous trip ended. Unfortunately, each of these 11 trips is the first in the dataset for that customer.

At this point, remain only two feasible options: (1) impute missing using an arbitrary position in a 2 000 meters radius[39] around the second GPS coordinates available for that trip; (2) remove the record and assume the trip started in the second GPS location; (3) remove the entire trip.

Both the first and second options are not robust and would probably cause some kind of issue during the features engineering stage. Considering these 11 trips are characterized by a very short total distance travelled, to solve the issue it is decided to remove the entire trips.

[38] This means they could erroneously be counted as behaviour event whereas they are simply noise.

[39] Remembering it was explained the GPS position are usually recorded every 2 km.

Duplicated timestamp

The phenomenon of records with duplicated "timestamp" values within the same trip cannot be neglected because it affects 10.14% of trips and 0.61% of total records. The investigation of duplicated records starts from analysing timestamp because, theoretically, it is expected to have no more than one record per second. It is interesting to observe that this phenomenon concerns exclusively records related to *behaviour events*.

To solve this issue, there is a trade-off between not dropping duplicated records and losing records that seem duplicated but in reality, represent behavioural event of different nature recorded at the same time. It is worth mentioning that this trade-off is caused by the lack of the data source containing accelerometer information about each behavioural event, which precludes the project from interpreting the *behaviour events*.

It is discovered that 53.58% of these records are completely duplicated records, hence, are dropped. Then, considering the previously identified issue of poor "gps_quality", it is appropriate to remove those records with duplicated timestamp and duplicated "distance_elapsed"[40]. These cases accounts for a portion of 18.54%.

Additionally, 3.97% of the records with duplicated "timestamp" have duplicated "latitude" and "longitude", and "speed" equal to zero; they are removed. Considering the previously mentioned trade-off, it is decided to not remove the remaining 23.91% records with duplicated "timestamp" because they do not present duplicated values for any other variable.

speed

The main issue about this variable is the abnormal frequency peak in zero. The frequency of the peak is 2 667 449 records. Considering there is at least one record per trip featured with speed equal to zero, there must be a complementary explanation for the remaining 14.26% of records with "speed" equal to 0 km/h.

Another issue to analyse is the maximum value is 288 km/h which needs to be investigated for discovering if it is a measurement error or it is in the context of a very dangerous trip. The hypothesis is that a value of 0 km/h of variable "speed" represents in some cases the absence of kinetic motion and in others a Missing Not At Random.

The boxplot in Figure 3.11 aims to provide evidence for better understanding whether a causation relationship exist between "speed" equal to 0 km/h and poor levels of "gps_quality" – specifically, level 0 and level 1.

[40] Remembering that all records with "gps_quality" equal to level 1 are featured with "speed" equal to zero.

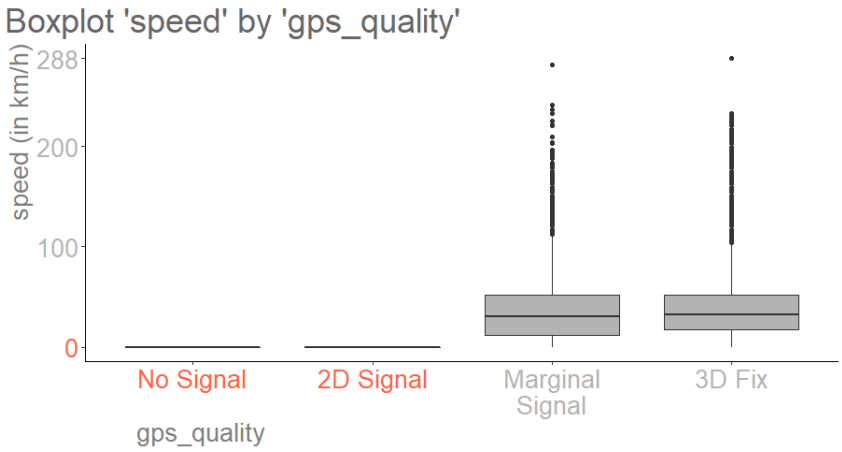


Figure 3.11 – Bivariate boxplot to investigate association between *speed* and *gps_quality*

Considering only the 2 667 449 (19.31%) records featured with "speed" equal to 0 km/h, 61.80% of them is also featured with poor levels of "gps_quality". Observing Figure 3.11 and considering variable "speed" has mean and variance equal to 0 for poor levels of variable "gps_quality", it is possible to conclude for those 61.80% of records – featured simultaneously with "speed" equal to 0 km/h and poor levels of "gps_quality" – variable "speed" represents a Missing Not At Random and not the absence of kinetic motion.

There are several possible alternatives to tackle this issue. One of them is to impute the missing values, for example by using an approach that combines rolling mean with logical rules. Another of the many possible approaches is to drop those records. For now, it is decided to not remove the 1 648 602 records with missing "speed". It is appropriate to investigate in-depth the effect of poor GPS quality on the other variables.

The conclusion of this investigation for variable "speed" is that it is more robust to find reliable alternatives to describe information about speed attributes of trips. For the clustering objective of this project, the solution is to compute the average speed using variables "distance_elapsed" and "time_elapsed". However, this information is only available for *position point* records.

heading

For variable "heading" it was observed a concentration of 1 271 623 records – 9.21% of the dataset – featured with variable "heading" equal to zero. It was already stated they represent missing values because "heading" should range from 1 to 360 degrees and not from 0 to 360 degrees.

The first hypothesis to validate is whether the peak of missing "heading" is caused by the 696 653 Engine On and 696 653 Engine Off events – one for each trip. In Figure 3.12, the 1 271 623 records missing "heading" are subdivided for each level of variable "session".

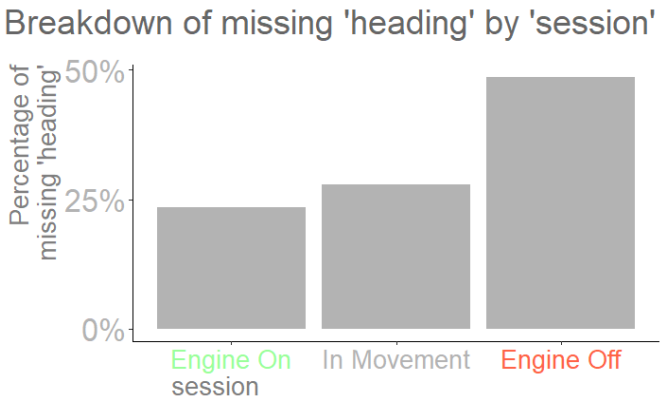


Figure 3.12 – Distribution total missing *heading* on categorical variable *session*

A considerable portion of the peak is caused by cars start up or turn off (72.14%). This confirms the expectation because telematics black-boxes computed variable "heading" using the motion of the vehicle: if the vehicle is moving, it is not possible for the device to compute the heading.

However, this insight is not enough to explain the peak because the remaining 27.86% missing values are generated while cars are travelling. Therefore, the investigation shifts toward understanding why "heading" is missing when the car is in movement session. The hypothesis is that poor levels of "gps_quality" impact on missing values of variable "heading" while cars are in movement "session".

In Figure 3.13, the bar chart aims to visualize the effect of variable "gps_quality" on missing values of variable "heading" when cars are in movement "session". It becomes clear that 55.57% of those 354 147 records missing "heading" while the car is in movement session are characterized by poor levels of "gps_quality". It is possible to conclude this is a causation relationship.

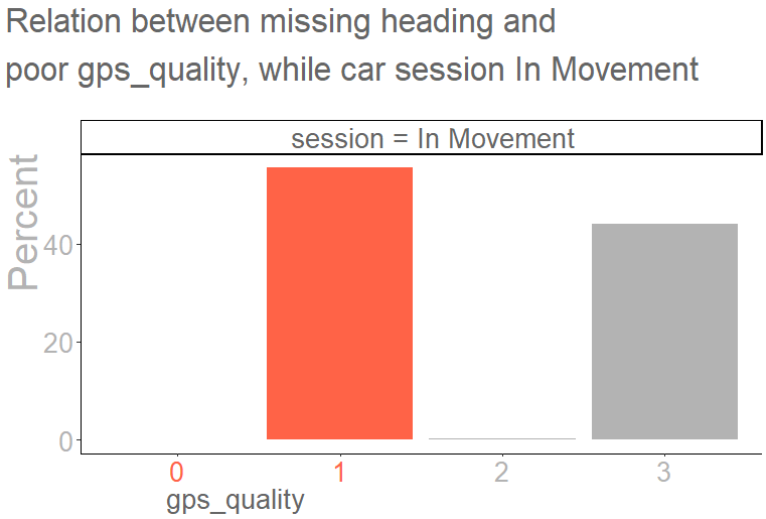


Figure 3.13 – Effects of *gps_quality* on records featured with missing *heading* and *session* in movement

After this second insight remains 44.43% of the records without clear explanation for missing values of variable "heading" – these are the records of Figure 3.13 featured with "gps_quality" level 3. In this case, the hypothesis to test is telematics black-boxes generate records with missing "heading" while cars are in movement "session" but the "speed" is equal to 0 km/h.

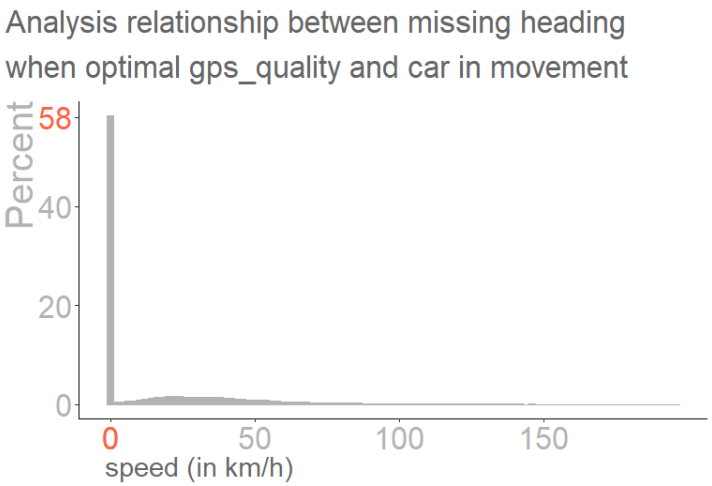


Figure 3.14 – Distribution variable *speed* for records missing *heading*, *gps_quality* level 3 and *session* in movement

Figure 3.14 shows the distribution of "speed" for records missing "heading" while the car is in movement "session" and the "gps_quality" is equal to level 3. This explains 58.30% of the remaining missing "heading" as caused by the absence of "speed".

At the end of this investigation, the 5.13% of records with missing "heading" remains with no explanation.

The main conclusion after this analysis is that using variable "heading" for creating summary variables to describe trips would require to: (1) impute missing values while cars are travelling and (2) neglect the "heading" of first and last record of every trip because are not meaningful. Additionally, the data collection sampling rate of 2 000 m is excessively large for computing meaningful variables based on "heading".

distance_elapsed

According to Section 3.2.3 "Raw Telematics Data Exploration", there are two main issues to investigate about this variable: (1) the high frequency peak in 0 meters; (2) values significantly diverging from 2 000 meters – which is the approximate sampling rate of data collection.

Without the data source related to behaviour events, it is not possible to reliably assess the quality of this variable. There are only two certainties:

- Records related to p*osition points* – with exception for Engine On – are featured with variable "distance_elapsed" higher than 0 meters, theoretically around 2 000 meters
- Records related to *behaviour events* are featured with "distance_elapsed" equal to 0 meters, representing a Missing Not At Random recoded as zero. In fact, *behaviour* events are not taken into consideration by telematics black-boxes for the computation of variable "distance_elapsed".

The Missing Not At Random are recoded as zero because variable "distance_elapsed" does not present any missing values but a 79.79% peak of records featured with "distance_elapsed" equal to 0 meters. Considering low levels of variable "gps_quality" have a significant impact on the overall quality of the records, it is reasonable to expect it to affect variable "distance_elapsed" as well.

Lacking the data source explaining the nature of each behaviour event, it requires to define assumptions on which records with "distance_elapsed" equal to zero are behaviour events and which are not.

The quality of raw telematics data has been investigated and assessed. Next, the raw data are enriched using external data sources freely available.

3.3.2. Data Enrichment

In this stage the initial raw dataset at the granular level of GPS coordinate points is integrated with geographical context gathered from open data sources and joined utilizing GIS methodologies. The integration of additional context involved combining datasets from multiple sources and different data formats. The objective is achieved by joining features from one spatial objects to another based on the spatial relationship.

In this section are presented only two of the many possible data enrichment for trips raw telematics data. Other potential data sources have not been integrated but should be added in future works; for example, information about vehicle characteristics or demographic profile of customers. More complicated example could concern social media or information about whether and road traffic conditions.

For the current study, the enrichment is related to geographical context and is based on spatial data combined with statistics about territories and information about roads network. The join operation for spatial data is technically referred to as spatial join. It differs from traditional join because it involves matching records based on their relative spatial locations.

The first geographical data enrichment task is related to data sources containing administrative information and national statistics about territories at the finest granularity level of municipalities. The second enrichment task is related to data sources containing information about roads network using OpenStreetMap[41] cartographies and data.

R Language comes at hand with useful frameworks and packages to handle spatial objects.

- The *sp* package contains spatial classes like SpatialPointsDataFrame, SpatialLinesDataFrame, SpatialPolygonsDataFrame, and others. Each of these classes consists of two objects: spatial object and structured metadata table. Depending on the class, the spatial object can be a set of points, a set of lines, a set of polygons, etc. Each set of spatial objects relates to a record in the structured metadata table.
- The *rgeos* package is an interface to the Geometry Engine – Open Source (GEOS) allowing to perform GIS techniques in R Environment using spatial classes contained in the *sp* package.

[41] Accessible through https://wiki.openstreetmap.org/wiki/Main_Page

- The *rgdal* package provides binding to the Geospatial Data Abstraction Layer allowing to read into *R* Environment popular geospatial vector format like shapefiles and to access projection and transformation operations[42].

The data used for enrichment are exclusively available for the Italian territory. The only information available for GPS coordinate points abroad is the fact they are not on the Italian territory. However, this is an acceptable compromise because only a marginal percentage (0.58%) of trips are abroad.

Source 1: Statistical and administrative information about municipalities

This task articulates over three steps. In the first step, descriptive statistics information is obtained by accessing open data available on the website of ISTAT[43]. This information is at the finest granular level of municipalities and is formatted in the structure shown in Table 3.4.

Field	Description	Note	Year	Source
Region ID	Region code in the range 01-20	Numeric	2017	Istat
Municipality ID	Municipality primary key	Numeric	2017	Istat
Name	Name of the municipality	String	2017	Istat
Altimetric Area	1= Inland Mountain; 2= Coastal Mountain; 3= Inland Hill; 4= Coastal Hill; 5= Plain	Categorical	2017	Istat
Altitude	Height above sea level of the municipality (in meters)	Numeric	2011	Istat
Coastal Municipality	1= Coastal, 0= Not coastal	Categorical	2017	Istat
Mountain Municipality	NM= Not Mountain Area, T= Totally Mountain Area, P= Partially Mountain Area	Categorical	1990	Uncem
Municipality surface area (in kmq)	The total surface of Italy is computed by adding up the surface area of each municipality.	Numeric	2011	Istat
Level of urbanisation	1= densely populated; 2= intermediate; 3= sparsely populated (rural area)	Categorical	2006	Eurostat

Table 3.4 – Summary of the table used for data enrichment resulting from the merge of multiple open data sources

The second step is fundamental because returns as output a spatial object of municipalities borders enriched with the information of Table 3.4 ready to be spatial joined with GPS coordinate points of raw trips telematics data.

First, the geospatial vector of Italian administrative boundaries at the granular level of municipalities is obtained from the open source "gadm.org". Function readOGR of *rgdal* packages allows to read the geospatial vector into *R* environment. Package *sp* allows to handle this data using the *SpatialPolygonsDataFrame* class[44]. The structured metadata table – describing each spatial object – contains five attributes describing each polygon: unique ID, country name, region name, province

[42] To operate with multiple spatial objects, they need to be projected on the same Coordinate Reference System.

[43] It is the Italian version of Instituto Nacional de Estatìstica of Portugal.

[44] SpatialPolygonsDataFrame class is composed of two data formats: (1) the projected GPS coordinates defining the border of each polygon; (2) a structured metadata table containing features describing the spatial object.

name and municipality name. The dataset is not officially produced nor updated by ISTAT, hence, the number of municipalities is 8100[45] instead of 7978.

Then, the join between the *SpatialPolygonsDataFrame* and the ISTAT dataset formatted as in Table 3.4 is performed. This join is accomplished by creating a composite key of region name and municipality name for both the data sources. Clearly, some data preparation is required to deal with all those issues arising when using strings of characters as joining key. The join based on string and the different number of municipalities of the ISTAT dataset compared to the geospatial vector obtained from "gadm.org" cause 3.56% of polygons to remain unmatched with ISTAT information.

The output of the second step is an enriched *SpatialPolygonsDataFrame* consisting of municipalities borders as spatial data and ISTAT statistical features as metadata. Finally, it is possible to perform the spatial join between GPS coordinate points of trips raw telematics data and the enriched *SpatialPolygonsDataFrame* containing ISTAT statistical and administrative information about municipalities.

In the last step the enrichment is accomplished by spatial joining trips GPS coordinates points with the *SpatialPolygonsDataFrame* polygons containing municipalities boundaries integrated with ISTAT statistical and administrative data. This functionality is offered by the function *over* contained in *rgeos* package. The solution to spatial join between points and polygons is simple although it is critical to project the spatial objects on the same Coordinate Reference System using *spTransform* function of *rgdal* package.

The enrichment of trips raw telematics data has a good quality considering 0.6% of the records are abroad of Italian territory and 1.5% of records belongs to the 3.56% polygons for which it was not possible to integrate with ISTAT data.

The following issues have remained unaddressed:

- Improve the match accuracy by solving minor issues related to match based on strings;
- Merge the small municipalities of GADM in order to be perfectly consistent and updated as ISTAT data;
- Manage the enclaved microstates of San Marino and Vatican City by considering them as additional municipalities of the region surrounding each of them.

Source 2: Roads network information

The objective of the second data enrichment task is to integrate trips with information related to roads network type – e.g. highway, secondary roads, living street roads, etc. This information is crucial because enriches the dataset with context about surrounding environment of trips and speed limit. A detail explanation of OpenStreetMap road types classification is presented in Appendix 1.

The task is difficult to complete because to assign each GPS coordinates point its road type – through reverse-geocoding – requires computation of geodesic distance between each GPS coordinates to all possible roads in the network. Each GPS coordinate is assigned to the segment of road minimizing

[45] Since 2012, very small municipalities have been aggregated to reduce expenses of public administration and gadm SpatialPolygonsDataFrame does not contains these updates.

the geodesic distance. Spatial joining between points and line is more complex than spatial joining between points and polygons because a GPS coordinate can only lay inside or outside a polygon.

There is a bottleneck caused by computational complexity. First, the amount of memory required is a matrix *P* x *R* where *P* is the total number of GPS points to enrich and *R* is the total number of roads in the network. Second, the amount of computing power to assign the road to one GPS point is determined by the total number of roads involved in the computation of each single point.[46] In the next subsection it is explained how the issue is addressed by designing and implementing a tailored *divide-and-conquer* algorithm exploiting multicore processor.

The road network information is stored in ESRI shapefiles obtained from OpenStreetMap databases available for free download[47]. It is possible to access data sources at different granular levels, from macro areas composed of several countries to individual regions of a specific country. It was decided to use the finest granular level because it guarantees a very detailed road network. The ESRI shapefiles are composed of multiple layers but exclusively the *roads* layer is considered for the current study.

The road network at the granular level of entire Italy is stored in one ESRI shapefile and contains approximately $1 \cdot 10^6$ roads. The road network at the finest granular is stored in five ESRI shapefile – North East, North West, Centre, South, and Islands – and contains approximately $5.8 \cdot 10^6$ roads. The latter is incredibly more detailed and drastically improves enrichment accuracy. On the other hand, the computational complexity increases in the same magnitude.

Reverse-geocoding solution

Solving this task exclusively using *R* is challenging because it is a typical task for a *GIS* software which is optimized by design for performing task involving spatial objects. The algorithm to solve this task is composed of three main steps: an initialization stage, an external cycle and an internal cycle.

There is one initialization stage which create the link between each GPS coordinates in the dataset and the geographical area it belongs to – North East, North West, Centre, South or Islands. This allow to consider only GPS coordinates laying in North East territories when spatial joining with roads network of North East territories. Source 1 (Statistical and administrative information about municipalities) assigned the corresponding region to each point. Using this information, it was straightforward to link each region to one of the five territories.

After the initialization stage there is the main cycle which is iterated for each of the five territories. The roads network layer stored in the ESRI shapefile is loaded into a *SpatialLinesDataFrame* class. Next, the *SpatialLinesDataFrame* is cleaned from pedestrian and bicycle paths. Each set of lines correspond to a road related to a record in the structured metadata table containing information about road name, road type and other detailed information.

To deal with the bottleneck issue previously explain, the solution implemented for this project is based on a divide-and-conquer approach. A grid composed of 400[48] equal-sized squares is overlaid

[46] For example, considering $3 \cdot 10^6$ GPS points and $1 \cdot 10^6$ roads, it would require storing in memory a matrix of $3 \cdot 10^6$ x $1 \cdot 10^6$ and it computing $3 \cdot 10^{12}$ distances.

[47] Source: http://download.geofabrik.de/

[48] This number was determined through a heuristic process.

both the *SpatialLinesDataFrame* and GPS coordinates of the current iteration. Two auxiliary tables are obtained: (1) "point_lookup" table which relates each GPS coordinate to the grid it belongs to, and (2) "line_lookup" table which relates each road the grid it belongs to.

The internal cycle is iterated for each of the 400 grids. On each iteration, it loads only GPS coordinate points and roads contained within its boundaries. Then, a spatial transformation is applied to project the two spatial objects on the same Coordinate Reference System. Finally, it is performed the spatial join between points and line, and each point is enriched with roads characteristics of the nearest line.

In case the current grid is composed of more than 800[49] GPS coordinates, the spatial join is performed through parallelization of the task by partitioning points based on their primary key. Through this approach, each of the four processor resolve the spatial join of 200 GPS coordinates points by working on the same *SpatialLinesDataFrame*.

The raw data at the granular level of GPS coordinate points have been enriched. In the next section, the format of the data is changed in order to obtain the structure of the final dataset to feed the clustering algorithm.

3.3.3. Data Formatting

The initial raw dataset has been validated and enriched with georeferenced data freely available. Before proceeding to data pre-processing and modelling phases, it is necessary to change structure of the initial raw dataset for two reasons: (1) obtain the intermediate dataset at the granular level of *segments*[50]; (2) make the dataset suitable for k-means algorithm which is not designed to deal with longitudinal data[51].

The intermediate dataset at the granular level of *segments* is obtained by aggregating the initial raw dataset into *segments* of consecutive *position points*. In the resulting structure, each record represents the *segment* of distance travelled between two consecutive *position points*. *Behaviour events* are excluded from the structure and are used to create a variable describing the frequency of *behaviour event* happened in that *segment* of trip. In Table 3.5 is shown the results of applying this process to the trip presented in Table 3.2. Note that values "point_A" and "point_B" are primary keys to lookup the information in the initial raw dataset at the granular level of GPS coordinates points. Variable "distance" is expressed in kilometres and "time" in hour so to simplify the computation of average speed in km/h.

[49] Through heuristic, it was determined that parallelization was more efficient in dealing with a greater number of points.

[50] As previously explained, the *segments* are portion of roads between two consecutive *position points*, for which the road distance travelled is known.

[51] The initial raw dataset is longitudinal data because tracks the sample object at different point in time.

trip_key	point_A	point_B	distance_elapsed	time_elapsed	n_behaviour	road_type
1	1	2	2.051	0.072	0	service
1	2	3	2.002	0.043	0	unclassified
1	3	4	2.032	0.046	0	primary
1	4	5	2.058	0.031	0	primary
1	5	6	2.006	0.049	0	primary
1	6	7	2.025	0.038	0	trunk
1	7	12	2.092	0.042	**4**	trunk
1	12	15	2.017	0.032	**2**	trunk_link
1	15	18	0.567	0.110	**2**	unclassified

Table 3.5 – Sample of the intermediate dataset at the granular level of *segments,* using the trip of Table 3.2 enriched with road type information

This new structure overcomes two issues: (1) lack of the data source containing accelerometer information describing the nature of each behaviour event; (2) missing values of "distance_elapsed" for records related to *behaviour events*, making it unreliable to compute the distance travelled since the previous geographical position. In fact, an alternative to this approach – of creating an intermediate dataset at the granular level of *segments* between consecutive *position points* – could be to estimate the distance travelled for each record related to behaviour events. A possible technique such estimation could be to use the great-circle distance, measuring the shortest distance between two points on the surface of a sphere. However, this is complex, not reliable nor robust, because of the road network settings and the short distance[52].

By aggregating the intermediate dataset at the granular level of *segments*, it is possible to obtain the structure of the final Analytical Base Table at the granular level of trips. According to this structure, the information contained in the initial raw dataset are aggregated and summarized into one single record describing each trip.

The structure of the final Analytical Base Table is presented in Table 3.6. Again, temporal and spatial variables are expressed respectively in kilometres and hours. Variable "n_GPS" represent the total number of *position points*, including Engine On and Engine Off events.

trip_key	distance_length	time_length	n_behaviour	n_GPS	start_lat	start_long	end_lat	end_long	start_timestamp
1	16.85	0.46	8	10	44.******	10.******	44.******	10.******	7/28/2017 2:28:43 PM

Table 3.6 – Sample of the structure of the Analytical Base Table at the granular level of trips, using the trip of Table 3.5

Finally, the dataset is ready for pre-processing and modelling phases. In the next section, new features are created for describing each trip, the resulting dataset is filtered from noise and treated from outlier, and, finally, it is obtained the final Analytical Base Table to feed the clustering algorithm.

[52] Considering the "distance_elapsed" for behavioural point is always less than 2 000 meters, the great-circle distance computed on short ranges can diverge significantly from the real road distance travelled by the car. This would reflects on unreliable values of average speed.

3.4. DATA PRE-PROCESSING

According to the theoretical framework of CRISP-DM methodology, the data pre-processing phase belongs to Data Preparation phase. In this project, they are separated because in Section 3.3 "Data Preparation" are solved all those tasks required to obtain the structure of the final data set starting from the initial raw dataset. Instead, in this section are included all tasks implemented to obtain the final Analytical Base Table to feed the clustering algorithm starting from the output of data preparation phase.

These tasks include creation of new features, filtering of noisy trips considered not relevant for the scope of this analysis, and treatment of outliers that would cause the k-means algorithm to converge on solutions very far from the global optimum. Finally, a brief multivariate exploratory analysis is presented to discover first insights and understandings about trips.

3.4.1. Features Engineering

One of the most important factor to determine success or fail in data science projects is the features used. Together with data preparation phase, this probably is where most of the effort are invested. This makes sense, especially considering how time-consuming are to gather data, integrate it, clean it and pre-process it, and how much trial and error can go into features design. Features engineering is the result of an iterative and cyclical process of running the model, analysing the results, modifying the data, and cyclically repeat using new insights discovered or lessons learned from mistakes. Additionally, features engineering is difficult because it is domain-specific of insurance business. These are some of the reasons why it required great efforts.

The features created in this project incorporate information from enriched initial raw dataset, intermediate dataset at the granular level of *segments* and final dataset at the granular level of trips. Over 100 variables are created including categorical variable obtained through data enrichment described in Paragraph 3.3.2 "Data Enrichment". While each of these variables is extensively studied to produce exploratory data analysis outputs for Sterling Insurance, only the most relevant are used for the clustering model.

To understand trips characteristics by insurance perspective, the most appropriate variables are grouped in four macro areas used by the business to study risk:

- As: refers to the amount of risk exposure like total amount of kilometres travelled or total amount of time driving in the trip. This area is self-explanatory.
- When: refers to the temporal dimension of risk exposure like day of the week or day time. For example, the highest number of fatality accidents happens of late night of Friday.
- Where: refers to the geographical dimension of risk exposure like road type or territory. For example, it was already explained Naples is characterized by higher risk.
- How: refers to the behaviour of customers and how it positively or negatively impacts on risk exposure. For example, the ratio of distance travelled at a speed above 100 km/h.

In Table 3.7 is presented the summary of features created to describe trips in the final dataset.

Category	Family	Description	Number of variables
As	Distance	Information related to trips length	6
	Behavioural Events[53]	Information related to frequency of behavioural events	8
	GPS Quality	Information related to GPS quality	3
	Parked Time	Amount of time elapsed since previous trip	1
When	Time	Information related to temporal dimensions of trips	28
Where	Coordinates	Information derived from latitude and longitude	6
	Administrative Area	Categorical information described in Paragraph 0	12
	Road Type	Categorical information described in Paragraph 0	14
How	Average speed	Information derived from speed	18
	Speed	Information derived from speed	9
	Turn	Information derived from heading	2
	Direction	Information derived from heading	10
Total number of features describing trips			117

Table 3.7 – Summary of features created for each macro category

After all features are created, it is performed an extensive exploratory data analysis of the dataset at the granular level of trips. The exploration is from both univariate and multivariate perspective and has two-fold objective: develop understanding and knowledge, and identify issues requiring to be addressed for obtaining the final Analytical Base Table to feed the clustering algorithm. The analysis is performed in-depth and allows to understand patterns and additional issues to deal with before the modelling phase. The current report already describes in detail a data analysis exploration in Section 3.2.3 “Raw Telematics Data Exploration”, therefore, it is not possible to excessively dwell in the description of this second exploratory analysis. The decision is to exclusively state in this section the most crucial insights that will drive the next steps.

Among the many, it is discovered an important insight: 39.68% of trips are composed of only 2 *position points*. This means the only information available are those related to the first and last GPS positions: start and end of the trip. Hence, according to the theoretical sampling rate defined by Octotelematics, the travelled distance of these trips is expected to not significantly exceeds 2000 meters. This has an important impact for two reasons:

- Shorter trips mean it is not possible to discriminate the trip other than using distance
- Considering they are composed of only one segment, it is not possible to compute standard deviation for variables.

Even though the exploratory data analysis of the final dataset does not have a standalone section in this report, its insights are included and described throughout the next pre-processing tasks, and they shape modelling decisions.

[53] It was decided to not include them in the How area because it is not possible to understand the nature of each behavioural event since the dataset is not available.

3.4.2. Noise Filtering

Before proceeding with coherence checking and outliers' treatment, it is appropriate to filter those trips that would introduce noise in the analysis. For the current project, noise refers to two aspects: (1) trips that are not relevant for the discovery of behavioural patterns; (2) trips which information cannot be completely reliable. The rules defined and their impact on the final dataset are summarized in Table 3.8.

Noise issue	Affected trips
Total distance less than 300 meters	46 649
At least half of points with poor GPS quality	36 778
Missing coordinates of Engine On event	11
Total trips removed	**74 048 (10.62%)**

Table 3.8 – Rules defined for noise filtering and number of trips removed

Excessively short trips

During the exploration of the final dataset at the granular level of trips it was discovered that 39.68% of trips are composed of only two *position points*: the Engine On event and the Engine Off event.

Considering the sampling rate of *position points* is approximately 2 000 meters, a great portion of the trips are short – because the overall distance is not more than ≈2 000 meters – and scarcity of samples – because all information available relates only to start and arrival GPS coordinates.

Clearly, these trips are very difficult to be discriminated within themselves because they are short and contain little information, making them hard to be characterized and very similar. At this point, it is expected there will be clusters significantly more populated than others reflecting this peculiarity of the dataset.

Trips are noise for the current analysis when they are so short to become irrelevant in the discovery of patterns. The decision is to noise filter trips featured with total distance lower than 300 meters. Logically, it would have been more appropriate to set a slightly higher threshold – e.g. 900 meters – however, 14.41% of trips in the dataset have a total distance between 300 and 900 meters.

Poor quality

Trips characterized by an overall poor quality are filtered as noise because information and characteristics describing them are not reliable.

For each trip is computed the relative frequency of records featured with poor levels of variable "gps_quality". Those with more than 50% of poor quality records are filtered.

The threshold does not include cases with exactly 50% of poor GPS quality because, otherwise, this threshold would have excluded an additional 9.56% of trips; this because 39.68% of trips is composed of just 2 *position points*.

Missing Engine On event coordinates
This issue was already identified on the raw data at the finest granular level of GPS coordinates in Section 3.3.1 “Data Quality Assessment”. It was already tried to solve the issue without succeeding.

The trips are entirely removed from the final dataset because their information is not reliable. Anyway, an alternative approach could have handled them as measurement error.

3.4.3. Coherence Checking

The features engineering stage added new variables to the structure of the final Analytical Base Table presented in Table 3.6. At this point, there are enough information available to start investigating the overall coherence of information contained in trips. The results summarized in this section have been identified through an iterative and cyclical process of investigation.

The objective of this stage is to discover whether there are trips featured with characteristics that are inconsistent among them. In this study, the variables used for this scope are:

- “n_GPS”: total number of *position points*[54]*;*
- “distance_length”: total distance travelled calculated as the sum of variable “distance_elapsed” described in Section 3.2.3 “Raw Telematics Data Exploration”;
- “detour”: ratio between total great circle distance[55] and variable “distance_length”;
- “delta_distance_parking”: great circle distance between arrival coordinates of previous trip and starting coordinates of next trip.

In the next subsections are described insights concerning coherence checking of variables “n_GPS”, “distance_length” and “detour”. Then, the decisions taken to deal with identified inconsistency are described in the subsection concluding the description of coherence checking task.

The insights obtained from coherence checking “delta_distance_parking” are out of scope of this project because it is discovered the telematics black-box of two drivers frequently do not record long trips. This can be caused by device issues or by potential vulnerabilities of telematics black-boxes that can be exploited to fraud insurance companies.

distance_length and position points
The coherence between number of *position points* and total distance travelled is checked from multiple perspectives. Interesting insights are obtained by comparing the number of *position points* to the ration between total distance travelled and maximum theoretical distance.

The maximum theoretical distance is computed as 2 000 meters multiplied by the total number of *position points* minus one – to exclude the one related to Engine On event which value of variable “distance_elapsed” which is always equal to 0 meters (refer back to the example of Table 3.2).

[54] Remembering that the distance between *position points* should not exceed on average 2 000 meters.

[55] Computed based on haversine formula using as parameters first and last coordinates of each trip. Variable “detour” is based on the idea of highlighting trips started and ended in the same GPS location.

Figure 3.15 aims to represents through a visual approach this articulated coherence validation. On the x-axis are represented the total number of *position points* per trip. On the y-axis is represented the ratio between total road distance and maximum theoretical maximum distance. The value of 100% for y-axis – marked with the horizontal red line – represents trips which total distance travelled is exactly equal to the maximum theoretical distance. All trips below the horizontal red line are not visualized in the plot because their total distance travelled is lower than their maximum theoretical distance. All trips marked with the black triangle are inconsistent because their value of variable "distance_length" exceeds the maximum theoretical distance of that specific trip. This inconsistency is observed for 6.30% of trips.

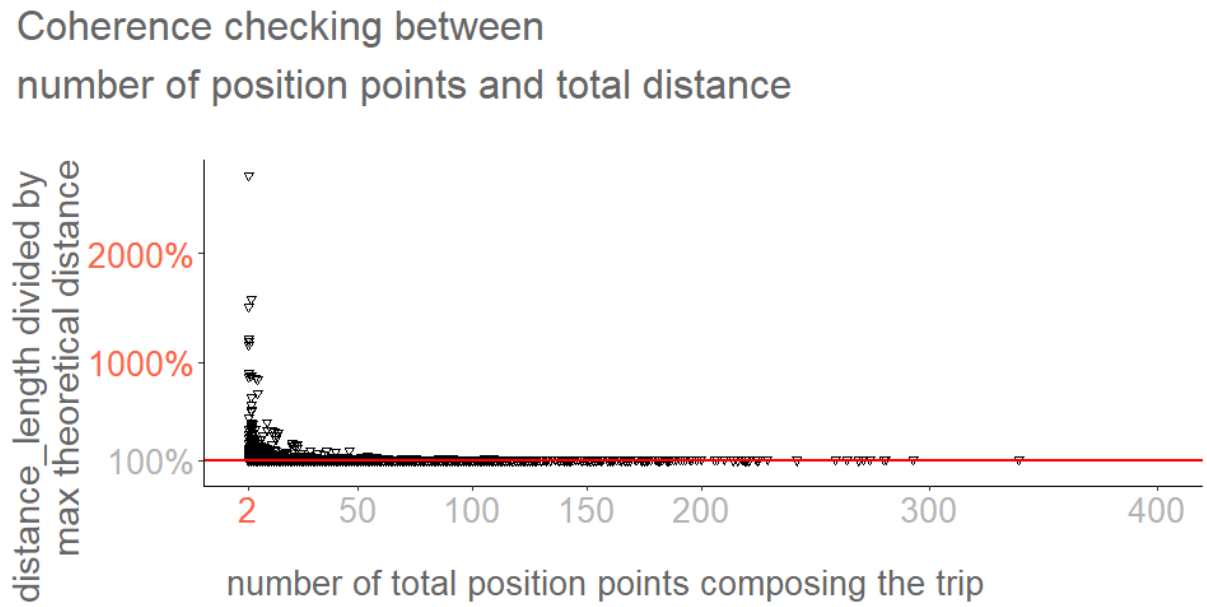


Figure 3.15 – Visual coherence checking of total *position points* and *total distance*

It is observed the inconsistency is significantly more relevant for trips featured with a small number of *position points*. The most inconsistent trip is composed of two *position points* and is featured with a value of "distance_length" 2714 times bigger than its maximum theoretical distance. In fact, its total distance travelled is 54 290 meters[56] and it far exceeds the approximate sampling rate of 2 000 meters.

For reasons yet to be discovered some trips are missing a considerable amount of *position points*.

detour

Variable "detour" represents the ratio between total great circle distance[57] and total road distance travelled. In the real world, total great circle distance can never be longer than total road distance.

Therefore, variable "detour" should theoretically range from 0 – representing the most indirect route possible: starting and ending in the same coordinates – to 1 – representing the most possible straight route between start and arrival coordinates.

[56] Considering it is composed of only one *position point* besides the other *position point* related to the Engine On event, its "distance_elapsed" is equal to the total distance travelled in the entire trip.

[57] Computed based on haversine formula using as parameters first and last coordinates of each trip.

The need of coherence checking variable “detour” was born after discovering outliers as a side-effect of running k-means algorithm on the final dataset at the granular level of trips. This investigation highlights different cases of inconsistency caused by telematics black-boxes measurement errors.

In Figure 3.16 is visualized the distribution of variable “detour”. It is immediately clear there is inconsistency because 3.48% of trips present an inconsistent value[58] of variable “detour”. To allow the visualization, all values above 1.9 ranging up to 1913.50 have been binned together.

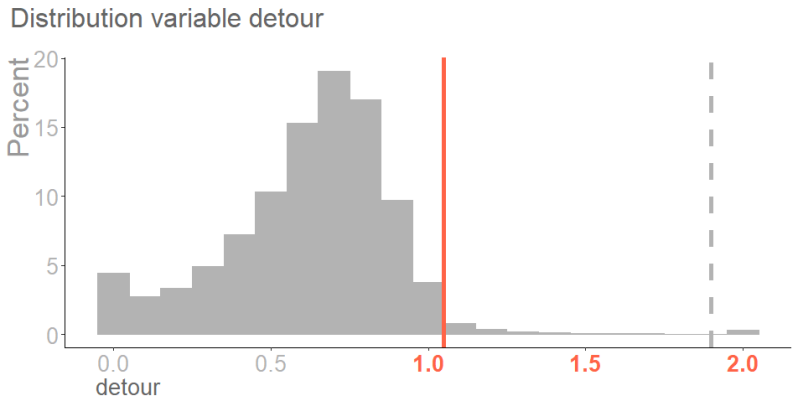


Figure 3.16 – Distribution variable “detour”

Considering only trips featured with an inconsistent detour value higher than 1 but lower than 1.9, 94.18% of them are trips with only two position points. The measurement accuracy error on variable “distance_elapsed” is more impactful on shorter trips, causing inconsistent values of variable “detour”.

Clearly, extreme values – for example the trip featured with a detour value of 1913.50 – require deeper investigation. This topic is discussed in the outliers’ treatment stage.

Decisions

There is a tight correlation between the coherence checking of variable “detour” and the coherence checking of distance travelled and number of *position points*.

In fact, 94.08% of trips characterized by inconsistent value of variable “detour” present a number of *position points* equal to two. This is explained by the fact that an inconsistent value of variable “detour” is expression of the inconsistency between “distance_length” and maximum theoretical distance.

It is possible to conclude there are two types of inconsistency both caused by telematics black-boxes measurement errors: (1) minor inconsistency caused by accuracy error in computing variable “distance_elapsed” at the granular level of GPS coordinates, and (2) major inconsistency caused by malfunctioning of unknown nature.

The trips affected by minor inconsistencies it is expected to not cause any serious problem in modelling phase. Instead, trips affected by major inconsistencies are extreme values that are addressed in the upcoming section.

[58] It is not possible for distance travelled to be shorter than great-circle distance between two points.

3.4.4. Outliers' Treatment

The decisions taken regarding how to treat outliers are the result of an iterative and cyclical process of modelling results comparison and investigation. Outliers treatment is as delicate as fundamental for this project, which clustering results rely on k-means algorithm. It is delicate because requires a meticulous study to discover phenomena behind extreme values for each variable. It is fundamental because k-means results are sensitive to outliers. Such sensitivity makes k-means an effective tool to detect outliers in the dataset.

Algorithms and automated methods for outliers' treatment exist, however, it is decided to approach this task as cautious as possible by manually treat each variable. Obviously, this was feasible because of the limited number of variables. The approach adopted supplied reasonable ground to exclude some trips from the learning stage. Once the clustering model is learnt, outliers are assigned to the cluster of the closest centroid[59].

In this project have been detected outliers of two different natures: (1) measurement error caused by telematics black-boxes malfunctioning[60]; (2) real observations that are both rare and far away from central tendency of the distribution.

In Figure 3.17 is visualized one example of outliers caused by telematics black-box malfunctioning. This is outlier related to variable "detour" and was discovered because k-means algorithm created a very small cluster including trips similar to this one.

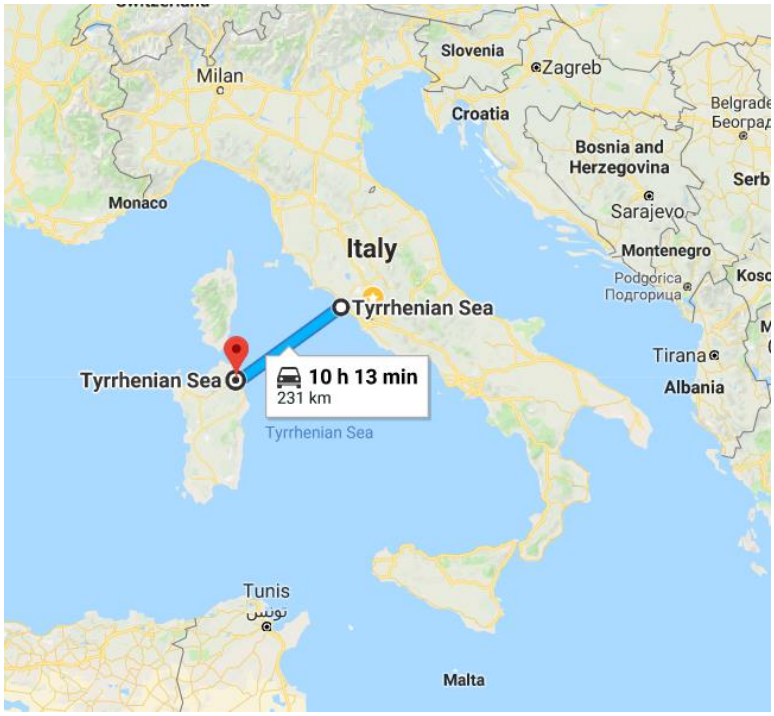


Figure 3.17 – Visualisation of a "detour" outlier caused by a ferry trip

[59] A more robust approach would have been to learn a multi-class predictive model over clustering results, and use the predictor to assign the outliers the cluster they belong to.

[60] Refers to major inconsistencies identified in Section 453.4.3 "Coherence Checking".

It is interesting to observe how telematics black-box misunderstood a ferry trip as a road trip. The resulting trip is composed of two *position points* featured with a road distance travelled of 253 meters and a total time of 11 hours. On the opposite, the great circle distance is 217 km, therefore, variable "detour" is 859.60; which is both a major inconsistency and an outlier.

Defining thresholds for outliers filtering requires to balance the trade-off between risk of learning a clustering model excluding important peculiarities of the dataset, and risk of learning a clustering model deeply influenced by characteristics that are not common neither generalizable to the entire population.

Variable	Description	Minimum	Maximum	Filter Method
distance_length	total distance		100 km	MANUAL
time_length	total time		120 MM	MANUAL
n_records	number of records		170	MANUAL
n_GPS	number of *position points*		55	MANUAL
n_behaviour	number of *behaviour events*		150	MANUAL
max_speed	maximum speed		250 km/h	MANUAL
avg_speed	average speed		150 km/h	MANUAL
detour	detour		1.5	MANUAL

Table 3.9 – Outliers filtering rules manually selected

Applying the outliers filtering threshold of Table 3.9, 2.48% of remaining[61] trips are excluded from the learning and will be assigned to clusters once the model will be built.

The output of the Data Pre-Processing phase is the final Analytical Base Table ready to feed the k-means clustering algorithm. This phase was crucial because better the quality of the data fed to the model, better the clustering results obtained. Summing up the effect of data preparation and data pre-processing phases, the

Task	Number of GPS points removed	Number of trips removed
Data Quality Assessment	1 081 036	NA
Noise Filtering	NA	74 048
Coherence Checking	NA	0
Outliers Treatment	NA	15 469[62]
Total	**1 081 036** (7.83%)	**89 517** (12.84%)

Table 3.10 – Recap of data cleansing effect on the dataset

[61] The outliers filtering is performed on the dataset resulting from noise filtering stage.

[62] In reality, the outliers are not removed from the dataset. They are not included in the learning of the clustering model but will be assigned to the cluster of their nearest centroid at a later time.

3.5. CLUSTERING MODEL

After several cyclical iteration of several CRISP-DM phases, it is obtained the final Analytical Base Table ready for the modelling phase.

Clustering is the data mining task of dividing individuals into groups of similar objects. Individuals groups together are similar within themselves and dissimilar between other groups. The representation of data using fewer clusters causes loss of fine details, but it achieves generalization and simplification. In machine learning, clusters represent hidden patterns in data. The search for clusters is unsupervised learning task, accomplished by clustering algorithms. In unsupervised learning, the training data used to learn the model are without label. A label is assigned by the clustering algorithm. Then, it is responsibility of the data scientist to describe typical behaviour of each group; this task is completely subjective and domain-specific.

The modelling phase includes tasks such as dimensionality reduction and variable selection, selection of the clustering technique, and clusters profiling to assess modelling results and improve approaches to modelling. This section concludes by explaining functioning and assumptions of k-means algorithm. The final model resulting from the current study is discussed in Chapter 4 "Results and discussion".

There are two important characteristics of the dataset to bear in mind during modelling phase. First, compared to the entire portfolio, this dataset is affected by seasonality because most trips were performed in summer period. For example, "distance_length" is 8.74% higher in summer trips compared to winter trips. Second, Octotelematics did not adopt any methodology to guarantee the dataset is representative of the entire population of 291 000 customers. Therefore, it is unlikely clustering results to be perfectly generalizable on the entire portfolio.

However, the current project defines a methodology that can be easily extended to the entire portfolio once the Big Data platforms will be implemented in operations.

3.5.1. Dimensionality Reduction

The decisions taken for dimensionality reduction are based on an iterative and cyclical process of modelling results comparison and investigation. While there are other methods, for example, Jebara & Jaakkola (2000), most of them are used primarily for supervised and not unsupervised learning, thus, they do not address general-purpose attribute selection for clustering. It can be concluded that cluster-specific[63, 64] attribute selection methods have yet to be invented.

In this project, dimensionality reduction is strictly related to variable selection and they are both critical tasks for the success of the current study. It was previously explained that *k-means* algorithm is a search strategy to minimize the objective function represented by sum of square over Euclidean distance. Considering this and considering 127 features were obtained in Section 3.4.1 "Features Engineering", it is worth thinking how Euclidean distance relates to the curse of dimensionality (Bellman, 1961).

[63] The computing resources and time constraints did not make possible to develop any calculation with memory complexity of $O(n^2)$ because it would have meant dealing with a table of size $49 \cdot e^{10}$.

[64] An additional complication is that the clustering tasks are highly subjective.

The short answer is that volume of the search space increases at an incredible rate relative to the number of dimensions. Even 10 dimensions – which might not seem very "high dimensional" – can potentially bring on the curse. If the data are distributed uniformly in that space, all objects become approximately equidistant from each other. However, there is another reason why dimensionality reduction and variable selection are crucial for this project. In fact, assuming a clustering model resulting from a *k-mean*s algorithm in which 98 out of 100 features are completely irrelevant for the business goal, their noise would completely swamp the signal of the two relevant features.

All features have been studied and their impact on clustering results extensively studied. From this iterative and cyclical process of trial and evaluation, it was gradually concluded to exclude all categorical features, exclude features related to "when" dimension and, concerning "where" dimension, to include only features related to road type and speed limits information.

Categorical features

In the final Analytical Base Table, the number of features to potentially use for clustering is 127. The categorical features are 20 – including 8 features derived from Source 1 (Source 1: Statistical and administrative information about municipalities). Theoretically, categorical variables are not suitable for algorithms which distance is computed in Euclidean space; this is the case of k-means algorithm chosen for this project.

To overcome this limitation there are several different approaches. For example, to change algorithm, to change distance metrics and space, or to apply variables transformation techniques. In this case, the transformation has the objective of transforming the data to fit k-means algorithm[65].

For operative reasons and time constraints, it was only studied the effect of one-hot encoding transformation technique on categorical variables. This transformation is very straightforward, creating *L* dummy variables where *L* is the number of different levels for the categorical variable.

However, in the case of the objectives of this project, the effect on clustering results was not satisfactory because this approach resulted in assigning very similar trips to different clusters solely based on the one-hot encoded categorical variables. This cannot be acceptable if the one-hot encoded categorical variable is not crucial from business perspective.

The categorical variables excluded in this stage might be useful for an additional layer of analysis after the clustering results has been deployed to profile customers. For example, variables related to altitude or time slot could potentially be used to further investigate customers characteristics of specific clusters. However, this is out of scope for the current project which objective is to study trips characteristics regardless of information strictly related to customers.

***When* features**

Features related to temporal dimension have been extensively investigated. After careful modelling iterations and meticulous study, it was decided to not include them in the learning of the clustering model because not relevant for the results.

[65] This is not a best practice because it should be the algorithm to fit the problem, not the other way.

Like what already observed in previous paragraphs, also these features are excluded from the final model because their only impact was to assign similar trips to different cluster solely based on the value of these variables. Again, this cannot be acceptable if the variable is not crucial from business perspective.

This is the case of all temporal features like “time_slot”, “day_of_the_week”, “month”, etc. An additional peculiarity of these feature is that they are cyclical.

***Where* features**

In section 3.3.2 “Data Enrichment” it was explained the methodology to enrich trips with geographic context using open data.

Most of the features related to statistical and administrative information about municipalities are categorical, hence, required to be one-hot encoded to be used in the k-means algorithm. These variables were analysed and extensively experimented. Again, they resulted in the assignment of trips having very similar characteristics to different clusters exclusively based on these variables. Again, this cannot be acceptable if the variable is not crucial from business perspective.

On the opposite, features related to road network information are extremely useful from business perspective. They allow to understand in which road context the trips were performed – road type – as well as to compare speed of vehicles to speed limit on that road.

Selected features

Before proceeding, it is important to remark that those features excluded are not relevant in discovering patterns over trips characteristics that are useful from business perspective.

The dimensionality reduction methodology adopted for features selection is based on the trade-off among three parameters: business interpretability, variables redundancy and relevancy of the results obtained from business perspective.

Business interpretability refers to the fact features need to be consistent with domain knowledge and meaningful for the company. Redundancy is evaluated studying the correlation between variables. Relevancy from business perspective refers to the ability of variable to influence k-means results by creating clusters that are more meaningful for the business.

After copious iteration and experimentations, the variables selected for the final clustering model are presented and briefly described in Table 3.11. This set of variables selected have been discovered as the most relevant and most meaningful from business perspective.

Name	Description	Datatype
DISTANCE_30_50_avg_speed	Percentage of distance travelled at an average speed between 30 and 50 km/h	Numeric
DISTANCE_50_90_avg_speed	Percentage of distance travelled at an average speed between 50 and 90 km/h	Numeric
DISTANCE_90_130_avg_speed	Percentage of distance travelled at an average speed between 90 and 130 km/h	Numeric
DISTANCE_130_000_avg_speed	Percentage of distance travelled at an average speed above 130 km/h	Numeric
n_GPS	Total number of *position points* represents a proxy for the total distance	Numeric
n_behaviour	Total number of *behaviour events*	Numeric
over_speed	Percentage of distance travelled exceeding the speed limits	Numeric
living_street	Percentage of distance travelled on "living_street" roads	Numeric
motorway	Percentage of distance travelled on "motorway" roads	Numeric
primary	Percentage of distance travelled on "primary" roads	Numeric
trunk	Percentage of distance travelled on "trunk" roads	Numeric
unclassified	Percentage of distance travelled on "unclassified" roads	Numeric
tertiary	Percentage of distance travelled on "tertiary" roads	Numeric
secondary	Percentage of distance travelled on "secondary" roads	Numeric

Table 3.11 – Brief description of variables used for final clustering model

It is particularly interesting that several variables which were expected to be relevant – like time slot in which the trip was performed – have not been selected since did not have useful impact on clustering results. Other variables – like total time length or total distance travelled – were excluded because perfectly correlated with others.

In Table 3.12 it can be observed the variables do not present significant linear correlation, making them not redundant.

	DISTANCE_30_50_avg_speed	DISTANCE_50_90_avg_speed	DISTANCE_90_130_avg_speed	DISTANCE_130_000_avg_speed	n_GPS	n_behaviour	over_speed	living_street	motorway	primary	trunk	unclassified	tertiary	secondary
DISTANCE_30_50_avg_speed														
DISTANCE_50_90_avg_speed	0.06													
DISTANCE_90_130_avg_speed	-0.04	0.18												
DISTANCE_130_000_avg_speed	-0.02	0.03	0.20											
n_GPS	0.18	0.51	0.48	0.20										
n_behaviour	0.23	0.32	0.17	0.10	0.40									
over_speed	0.13	0.49	0.30	0.24	0.29	0.22								
living_street	-0.13	-0.16	-0.08	-0.03	-0.15	-0.09	0.00							
motorway	-0.01	0.16	0.45	0.22	0.41	0.10	0.08	-0.06						
primary	0.10	0.19	0.02	0.00	0.15	0.08	-0.03	-0.13	-0.03					
trunk	0.01	0.28	0.46	0.14	0.31	0.10	0.20	-0.07	0.01	-0.03				
unclassified	-0.06	-0.15	-0.10	-0.03	-0.16	-0.08	-0.07	-0.21	-0.08	-0.17	-0.09			
tertiary	0.01	-0.10	-0.11	-0.04	-0.11	-0.03	0.05	-0.23	-0.09	-0.21	-0.11	-0.29		
secondary	0.07	0.06	-0.06	-0.02	0.01	0.05	-0.09	-0.22	-0.07	-0.20	-0.09	-0.27	-0.34	

Table 3.12 – Correlation Matrix using Pearson's Index for variables selected for final clustering model

For completeness of the study, in Appendix 3 is presented the Correlation Matrix computed using Spearman's Index.

3.5.2. Features Transformation

In the current project, features transformation accomplishes the goal of adapting features to k-means algorithm. All features are scaled through standardization. This is important to ensure that features with higher variance do not completely influence the final result of this study.

Additionally, in this project it was experimented a sinusoidal transformation applied to cyclical features like day of the week and start or arrival time. An appropriate approach should take into consideration the cyclical nature of those features. The concept is that cyclical features must be intended as circles, not straight lines. Dealing with daytime as linear would mean that 23 PM is the farthest point from 1 AM, whereas, in reality it is true the opposite.

The sinusoidal transformation used in this project consists in decomposing cyclical features in two variables that swing back and forth out of sync. Taking for example variable "start_time" – representing the time when trip started – the formulas used are the following:

$$hourfloat = \frac{start\_hour\ + start\_minute}{60}; \quad x = sin\ (\frac{\pi \cdot hourfloat}{24}); \quad y = cos\ (\frac{\pi \cdot hourfloat}{24})$$

The denominator is 24 because the day has 24 hours[66]. These two variables create a 24-hour clock and their combination univocally refer to a unique point in time as shown in Table 3.13.

hour	minute	x	y
0	00	0.00	+ 1.00
12	00	0.00	- 1.00

Table 3.13 – Each combination refers to a unique point in time

The next step consists of running a k-means algorithm on a sample composed of 100 trips, using only the two variables just created.

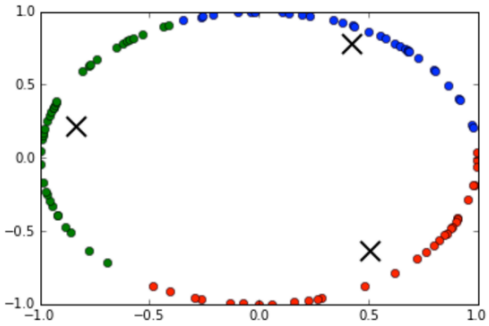


Figure 3.18 – Example of sinusoidal transformation on variable "start_time" cluster cyclical features

In Figure 3.18 it is visualized the 24-hour clock generated by the combination of the two variables. It is interesting to observe that k-means algorithm takes into account the cyclical nature of variable "start_time".

Unfortunately, when this methodology is studied on the entire dataset along with other variables, the result is not satisfactory because similar trips are assigned to different clusters solely based on these two variables. It is decided to exclude these variables from the final clustering model because their result is not relevant from business perspective.

3.5.3. K-means Algorithm

After dimensionality reduction and features transformation, the k-means algorithm to learn the clustering model is fed with the final Analytical Base Table. The project is based on k-means algorithms because its efficient in both memory usage and computational time, as well as because time constraints not allowed to experiment other clustering algorithms.

[66] It can be easily applied to day of the week by using 7 as denominator.

The k-means algorithm [Hartigan 1975; Hartigan & Wong 1979] is by far one of the most popular clustering tools. This algorithm has high notoriety even to non-technical, therefore, it was a safe choice. Additionally, it is not expensive in terms of both memory and computation time making it affordable to run several times on a large dataset[67] using different features and algorithm's parameters. The algorithm is efficient because it only computes distances between individuals and the *k* centroids. The downside of *k-means* is that it unlikely works well with categorical features. The name "k-means" derives from representing each of *k* clusters C_i by the mean c_i of its points, the so-called *centroid*. The objective function to minimize is the sum of dissimilarities of each group of objects with its centroid.

There are several parameters to set for this algorithm. (1) distance metric to use; (2) maximum number of iteration; (3) convergence condition; (4) number of random initial sets to evaluate; (5) number of clusters; (6) initial seeds.

The clustering task was approached from several different perspectives following a cyclical and iterative approach. Initially, several settings of input data, features and algorithm's parameters were explored, and insights discovered were implemented to improve data pre-processing and modelling decisions. Iteration after iteration Sterling Insurance' understanding has increased, and the focus shifted toward more specific direction: relevant features became clear and meaningful clusters were identified.

The pseudo code of the k-means algorithm used is the Hartigan-Wong in the R base package implementation written in Fortran language is presented in Figure 3.19.

Input: *k* (the number of clusters),
D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose *k* objects from *D* as the initial cluster centers;
Repeat:
1. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
2. Update the cluster means, i.e., calculate the mean value of the objects for each cluster

Until no change;

Figure 3.19 – Pseudo Code of k-means algorithm implemented in R "base" package

Clustering results of several different combination of variables have been studied in-depth. For each cluster analysis, it was generated from 2 to 24 different clusters. K-means algorithm is sensitive to initial seeds; therefore, it was chosen to evaluate 25 different initial set of seeds using *nstart* parameter. The decision was arbitrary though under the constraints time and computing resources available.

[67] Large dataset makes not feasible to use algorithms which require to compute a dissimilarity matrix.

4. RESULTS AND DISCUSSION

All the insights discovered through trial and error have been reimplemented in a cyclical and iterative fashion to improve previous phases of the CRISP-DM process. It is not relevant for the scope of this report to describe clustering results other than the final model deployed in the business. In this chapter, the final clustering model is presented and its business application is discussed.

4.1. Optimal Number of Clusters

The first result to be discussed in this chapter is the number of clusters chosen for the final clustering model. The decision was driven from business perspective. There are several statistical indices to evaluate the cluster solution, but none of them was taken into consideration to choose the number of clusters. For completeness of the study, the result of a statistical index is presented in Appendix 4.

Initially, Sterling Insurance expected a number of clusters around 40. The implementation of a meticulous clusters profiling methodology allowed to discover business insights suggesting the most appropriate number of clusters to answer the business problem using the dataset available is 11. The decision represents the optimal trade-off between two fundamental parameters: descriptive power of clusters and univocal association between customer and cluster. In fact, more clusters allow better interpretability of each trip, but an excessive number of clusters makes impossible the univocal association customer/cluster.

This concept applied to the final clustering model is visualized in Figure 4.1. The visualization aims to capture the density of maximum representation between clusters and voucher. It represents the maximum percentage of clusters association for each voucher. The vertical red dashed line represents the minimum percentage for that number of clusters, e.g. for 2 clusters, the minimum is 50%.

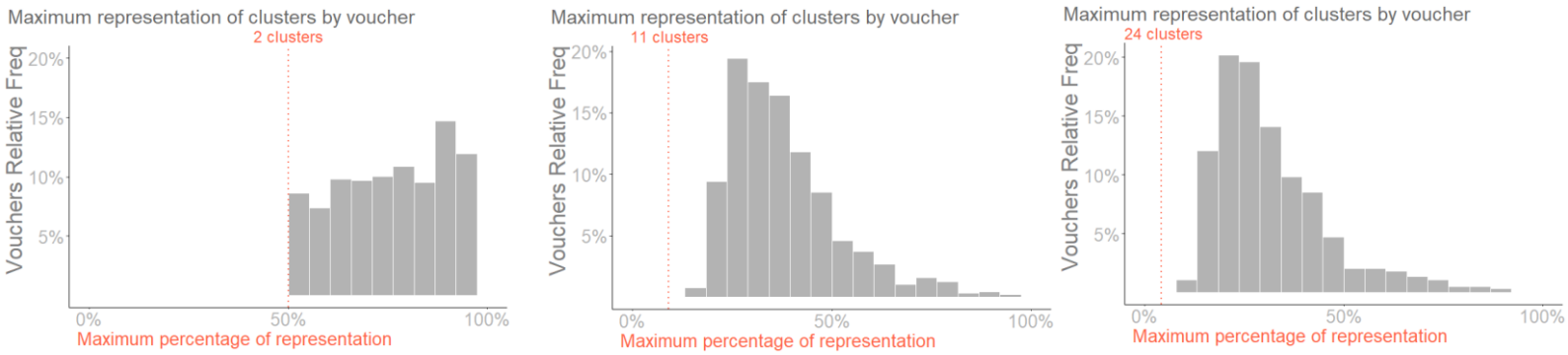


Figure 4.1 – Density of maximum representation of clusters by voucher

Using only two clusters guarantees the univocal association voucher/cluster because each customer can be perfectly profiled using no more than 2 clusters. On the other hand, the descriptive power of cluster would be extremely poor because one cluster – accounting for 77.52% of trips – is described as “very short distance travelled at low speed”, and the other cluster – accounting for 22.48% of trips – is described as “longer distance travelled at higher speed”.

Considering the opposite scenario with a number of cluster equal to 24, it guarantees very good clusters descriptive power, but the univocal association cluster/voucher is poor because each cluster is profiled using a high number of clusters.

The number of clusters optimizing association voucher/cluster and clusters descriptive power is 11: for 86.5% of customers, three clusters are enough to segment 60% of trips. It can be concluded the clustering model is robust enough from business perspective.

Concluding this section, it is important to remark this number of clusters is optimal considering the available dataset. In Section 3.2.1 "Data Source", it was discussed the absence for the current study of the data source containing accelerometer data describing each *behaviour event*. The impact was that it was not possible to associate the semantic meaning to any of the *behaviour events*[68]. It is reasonably expected the availability of such data source will significantly increase information available describing each trip and, therefore, reveal new patterns.

4.2. CLUSTERS PROFILING

The k-means algorithm allows to represent each cluster using its centroid. The centroid is the mean vector representing each cluster.

In Figure 4.2 are visualized the centroids of the 11 clusters. This approach allows to briefly grasp the broad outlines of each cluster. In reality, a tailor-made algorithm for clusters profiling was developed in order to allow a sound and deep comprehension of each cluster. Detailed visualization for clusters profiling are presented in Appendix 5. The description of each cluster is presented in Table 4.1.

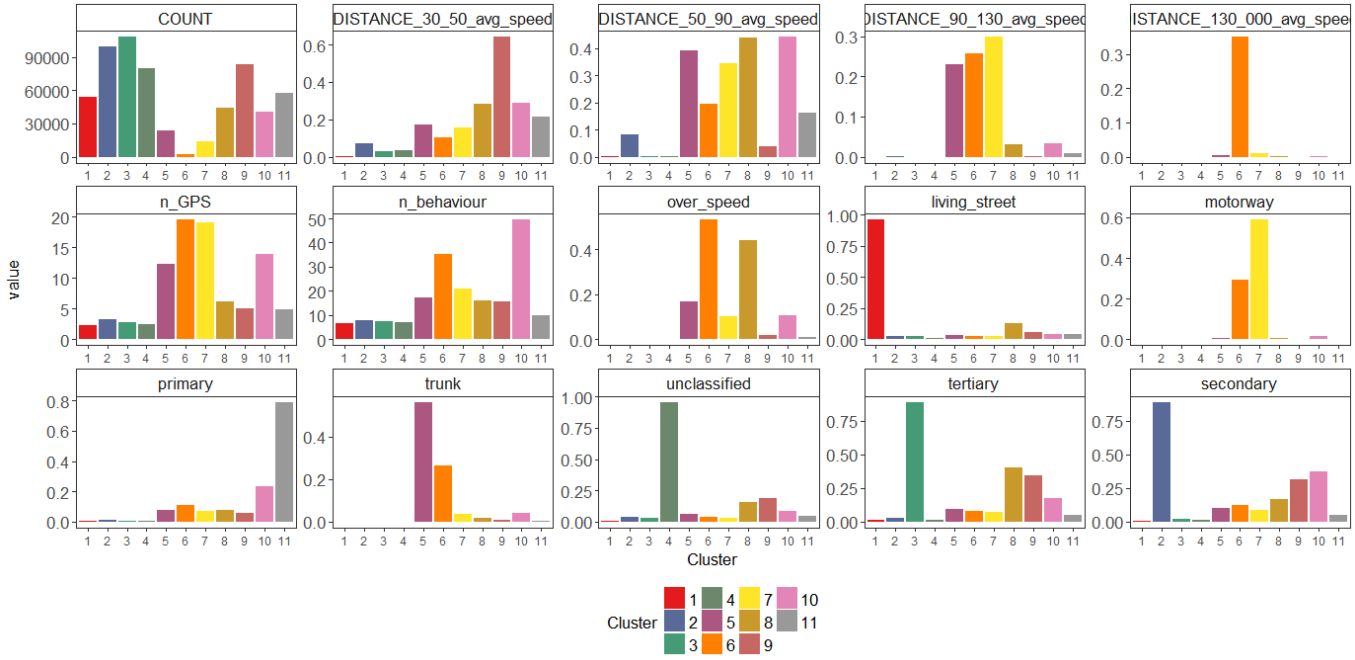


Figure 4.2 – Visualization of centroids for each cluster

[68] As studied in Section 3.2.3 "Raw Telematics Data Exploration" the number of *behaviour events* account approximately for 74.72% of the entire raw dataset at the granular level of GPS coordiantes.

In Figure 4.3 is visualized clusters distribution of trips for Sterling Insurance connected car portfolio. In red are highlighted clusters characterized by very risky behaviour. Cluster 10 is coloured less bright because its only dangerous characteristics is a slightly over speeding tendency at moderate speed.

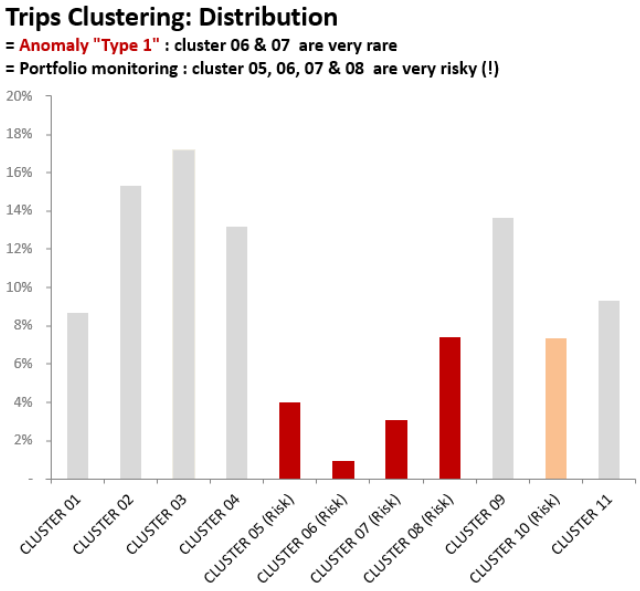


Figure 4.3 – Global distribution of trips by cluster (all portfolio drivers)

In Table 4.1 are presented description, frequency and risk associated to each of the 11 clusters obtained. The description is based on the study of several outputs in the form of visualizations and statistical analysis. The risk label is assigned through domain knowledge process supported by the business experts of Sterling Insurance.

Cluster_ID	Description	Frequency	Risk
01	Ordinary, very short trips exclusively on “living streets”	8.89%	
02	Ordinary, short trips exclusively on "secondary" roads	16.40%	
03	Ordinary, very short trips exclusively on “tertiary” roads	17.77%	
04	Ordinary, very short trips exclusively on “unclassified” roads	13.19%	
05	Long, fast trips on “trunk” roads	3.90%	**Yes**
06	Extremely fast, long trips on “motorway” roads	0.40%	**High**
07	Long, fast trips on “motorway” roads	2.35%	**High**
08	Over speeding trips on “tertiary”, “secondary” and “unclassified” roads	7.24%	**Yes**
09	Ordinary trips on “tertiary”, secondary” and “unclassified” roads	13.67%	
10	Slightly over speeding trips with frequent behavioural events on “primary”, “secondary” and “tertiary” roads	6.68%	**Low**
11	Ordinary trips exclusively on “primary” roads	9.52%	

Table 4.1 – Results of clusters profiling assign a meaningful label to each cluster

Before proceeding to apply clustering results to customers, the 15 469 outliers are assigned to the nearest centroid. The Euclidean distance to each centroid is computed and they are assigned to the one minimizing the distance. The result of this process is presented in Table 4.2:

Cluster	Percentage
1	8%
2	1%
3	1%
4	0%
5	1%
6	**21%**
7	**38%**
8	2%
9	1%
10	26%
11	1%

Table 4.2 – Summary outliers assigned to neared cluster

It is interesting to observe that approximately 21% of outliers are assigned to cluster 06 and 38% of outliers are assigned to Cluster 07, which are the riskiest and less populated clusters. This is evidence that a more robust approach was required. A more robust approach should have been to learn a multi-class predictive model on the clustered data to assign outliers the clusters they belong to. The advantage of this more robust approach would be that it discovers patterns of association and it considers variable importance in discriminating among clusters.

A further step in clusters profiling consists of identifying prototypes of each cluster, computed as the individual closest to the centroid. This operation is necessary in order to visualize on a map the trajectory of the most representative trip for each cluster. These visualizations confirm the descriptive power of clusters, and they are included in Appendix 6.

4.3. CUSTOMERS PROFILING

The characteristics of each cluster have been studied and a meaningful descriptive label has been assigned to each cluster. Finally, results of the study are deployed in the business by applying knowledge discovered through cluster analysis to profile all customers in the dataset. On the basis of these classes, each customer can be described according to the repartition of its trips across the 11 clusters.

Practically, this task consists of creating a relative frequency table in which each row refers to one customer and the 11 columns contain the percentage of trips made by that customer belonging to that specific cluster. Being relative frequency, each row sums up to 1 which corresponds to the total number of trips made by that customer.

It is important to remember that the clustering model was learnt on trips characteristics regardless of subjective information related to the customer whom performed those trips. This means the patterns discovered are highly generalizable and common to every customer.

A sample of the output produced is shown in Table 4.3. The lighter colour of Cluster 10 indicates its patterns are less risky compared to the others risky clusters.

VOUCHER	CLUSTER 01	CLUSTER 02	CLUSTER 03	CLUSTER 04	CLUSTER 05 (Risk)	CLUSTER 06 (Risk)	CLUSTER 07 (Risk)	CLUSTER 08 (Risk	CLUSTER 09	CLUSTER 10 (Risk)	CLUSTER 11	NUMBER of trips by voucher
3297776	23%	42%	2%	21%				2%	4%	2%	4%	2477
3803987	5%	18%	6%	6%	3%	1%	1%	14%	15%	26%	5%	2308
4586137	25%	18%	21%	11%	3%	0%	1%	4%	6%	6%	5%	2298
3181004	5%	9%	20%	23%	0%		1%	2%	29%	10%	1%	2158
3167949	3%	23%	5%	10%	6%	0%	1%	24%	19%	6%	4%	2104
4787692	17%	24%	25%	2%	4%	0%	1%	6%	17%	3%	1%	2097
3285341	11%	5%	18%	21%				3%	17%	0%	24%	2053
etc…												etc…
Average	**9.0%**	**15.0%**	**16.6%**	**12.1%**	**4.2%**	**1.2%**	**4.2%**	**7.3%**	**13.2%**	**7.6%**	**9.6%**	**663**

Table 4.3 – Sample of the final output knowledge for the 7 customers with more than 2 000 trips

This output is very relevant for the business because it is foundation for deploying the solution. This file is source of new, detailed risk knowledge about customers behaviour allowing Sterling Insurance to profile them from a new, detailed perspective.

4.3.1. Driver monitoring for dangerous customers

The most immediate use of this knowledge is to identify customers which majority of trips are dangerous.

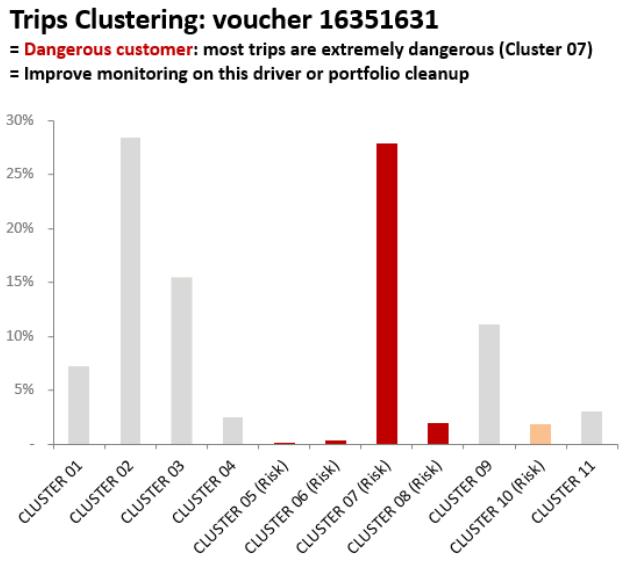


Figure 4.4 – Profile of dangerous customer

The insights of Figure 4.4 suggests to take action by monitoring or portfolio clean-up. A great percentage of trips performed by customers associated to profile similar to the one of voucher_id 16351631 are extremely dangerous. In fact, trips belonging to Cluster 07 are defined as very risky for Sterling Insurance.

4.3.2. Reward good customers

The knowledge derived from Table 4.3 allows to discover customers characterized by very safe profile. For example, the insights of Figure 4.5 suggests to improve the relationship with customers associated to profile like the one of voucher_id 18125420, which is characterized by a safe and careful driving behaviour.

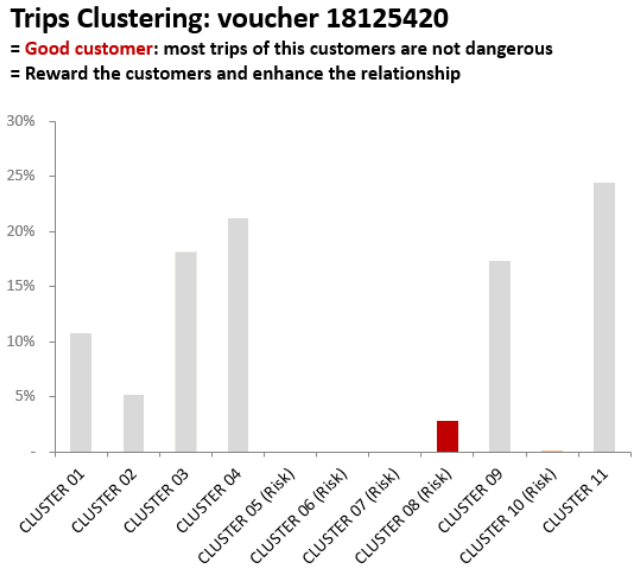


Figure 4.5 – Profile of good customer

4.3.3. Discover unusual trips

An additional example of possible business applications of the knowledge derived from Table 4.3 is the discovery of trips that significantly diverge from the typical behaviour of customers. This becomes especially true when trips made by the customers are concentrated in few clusters.

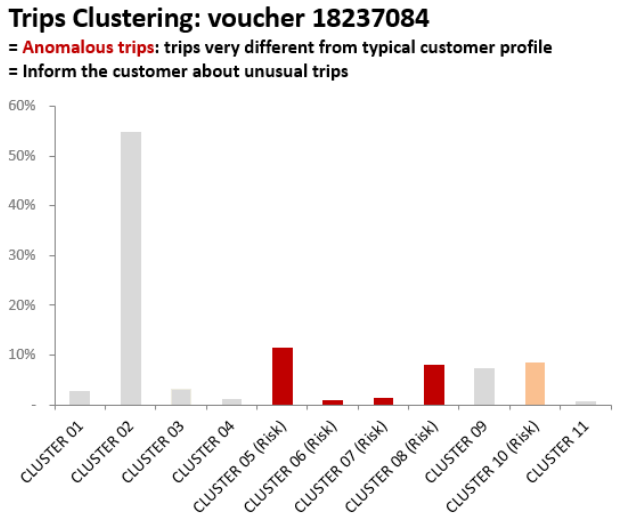


Figure 4.6 – Example of discovery of suspicious trips

In Figure 4.6 it can be observed that customers associated to profile similar to the one of voucher_id 18237084 mainly performs trips belonging to Cluster 02. For this specific example, suspicious trips are those related to Cluster 07 and 08 for two reasons: (1) the customer is characterized by a safe profile, and (2) the suspicious trips are related to very rare and dangerous clusters.

The insights of Figure 4.6 can suggest checking whether customers presenting this pattern are aware of those suspicious and dangerous trips. For example, it is common that people besides the policyholder make use of the car without him knowing.

5. CONCLUSIONS

Data science methods are new to many industries and they have the full potential to create new, disruptive business models. Firstly, as a general consideration, this project can also be solely intended as an example of how data science methods are generally applicable: knowledge extracted from data is something that can lead to improvements of operational processes and creation of new business models in all markets. According to this possible perspective, the contribution of this study is to propose an example of real value that can be extracted from data in any industrial sectors.

Considering car insurance market, in the past insurers could solely sell policies and pay claims because they did not have raw telematics data nor capabilities to extract knowledge from it. Today, we demonstrate Data Science methods applied to car insurance business allows to extract knowledge from data previously unavailable. This new knowledge becomes the foundation in revolutionizing customers relationship as well as in offering new services and products to market.

It is evident that research objectives and all specific objectives stated in the introduction were achieved. It was successfully designed and implemented data science methods and processes required to analyse and extract actionable knowledge from trips raw telematics data, in order to innovate understanding of risk knowledge associated to customers behaviour. Then, the new knowledge was used as foundation for new, innovative business routines aimed at taking different types of action over specific customers. Profiling customers based on their behaviour over each trip and using this knowledge to monitor and take action over dangerous customers or to reward good customers is something that could only be previously imagined. The results found are sound and pave the way for future large-scale projects.

We contributed in demonstrating that the use of raw telematics data innovates insurance telematics products and has the full potential to revolutionize Italian car insurance business over the next years. This study advances the state-of-the-art for vehicle telematics and it stands at the heart of the relevance for any new service based on vehicle telematics. The application of this research is mostly practical because the knowledge discovered is actionable in the sense that it can be directly applied by Sterling Insurance to take action – e.g. rewarding, portfolio clean-up, etc – over its customers. Additionally, the data science pipeline designed and implemented can be easily scaled up to the entire portfolio.

From the perspective specific to Sterling Insurance, it was achieved the milestone of developing skills and know-how related to vehicle telematics, essentials to design and offer new, high value-added services to both customers and companies. The strategic plan is aiming at building vehicle telematics expertise which is believed to be fundamental to gain competitive advantage against main competitors and future new entrants in the Italian car insurance market. While difficult to quantify, it is certain the organization has gain remarkable economic and knowledge benefits. The company is proprietary of the very first sound, in-depth data science research over raw telematics data, highlighting avenues for change and emphasizing the possibilities of evolution of insurer's offer, always more aligned with evolution of consumers.

To conclude, throughout this project I experienced the constant balance of the trade-off between operational needs and study completeness. I believe it is of critical importance for organizations to safeguard robustness and applicability of data science projects by approaching them with scientific

rigour, using an iterative and cyclical methodology aiming at converging for successive approximation. Data preparation, analysis and results interpretation require not only mathematical and statistical competences but also knowledge of the company and understanding in economics spheres including marketing. Rigidity in thoughts becomes obstacle that organizations need to overcome by creating teams composed of members from different backgrounds and multidisciplinary competences. The scope of data science projects must always be kept on the business value that can be created from data, and communication of results becomes as crucial as the study itself.

5.1. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Despite the very important results obtained, this study is subject to limitations in terms of data, computational resources available and time constraints. Data limitations are related to two aspects: lack of several data sources and low sampling rate. Computational resources and time limitations relates to compulsory decisions under way to simplify and respect the deadline for this project.

The most impactful limitation for this research is the lack of the data source containing accelerometer data related to *behaviour events* – which represent around 79% records of the initial raw dataset. This information would have allowed to assign a semantic meaning to each *behaviour events*. There is no doubt that including them would have greatly benefit and extended the results for this project. Additional data sources not used are those related to customers information and vehicle characteristics as well as external data sources such those related to weather or traffic conditions.

The low sampling rate – characterizing data used for this project – becomes a problem because 44% of trips are shorter than 4km and the low sampling rate reflects in the fact those trips are composed by only two points: Engine On and Engine Off. This makes it difficult to effectively discriminate those trips among themselves because very little information is available. In practice, low sampling rate causes considerable loss of information related to two dimensions: 1) trips summary characteristics are less detailed because are based on a very small sample of data, and 2) the trajectory of the trips is difficult to study especially in metropolitan area where there are many different combinations of streets connecting points that are 2 kilometres far from each other. The latter aspect do not impact this project but it will be an important issue to consider when performing data mining over the trajectory.

The limitations in terms of computational resources available and time constraints impacted on compulsory choices mainly limiting the choice of the clustering algorithm to adopt. Most of the efforts invested in the modelling phase were spent in dimensionality reduction and clusters profiling. There were not enough resources to experiment algorithms requiring an exponential complexity in terms of memory usage. There was not enough time to experiment more complex and ambitious clustering approaches.

The recommendations for future works that have highest priority for the Actuarial Department of Sterling Insurance are related to the implementation in the study of the lacking data sources: *behaviour events* data source and data sources related to customers personal information and vehicle characteristics. Additional data sources to implement with lower priority are those related to traffic or weather conditions.

From an academic perspective, it would be extremely interesting to deepen the study with more experimentation related to clustering algorithms and modelling approaches. Specifically, variables transformations could become more complex and clustering algorithms could be designed and developed tailor-made for trips raw telematics data.

From the business perspective of Sterling Insurance, the first step is to apply the model obtained from this study to the entire portfolio of trips. The clustering model is expected to generalize to the entire portfolio of telematics customers. Next, raw telematics data related to cars crashes should be studied in order to discover different patterns related to claims event. Finally, the study of trips and crashes raw telematics data should be combined in order to discover trips patterns predicting crashes.

To conclude, considering the current project performed clustering over summary characteristics describing each trip, an additional future recommendation is to perform data mining over the trajectory of trips. This will require higher sampling rate but would unlock an additional dimension of knowledge from trips raw telematics.

BIBLIOGRAPHY

ANIA. (2014). *Black-Boxes: Italy Global Leader*.

ANIA. (2016). *Italian Insurance 2015 - 2016*.

ANIA. (2017a). *Italian Insurance 2016 - 2017*.

ANIA. (2017b). *Italian Insurance in Figures*.

Baddeley, A. (2008). Analysing spatial point patterns in R. *Workshop Notes*, *12*(6), 1–199. https://doi.org/10.1007/s00415-011-6369-2

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, *98*, 69–79. https://doi.org/10.1016/j.dss.2017.04.009

Bellman, R. (1961). On the reduction of dimensionality for classes of dynamic programming processes. *Journal of Mathematical Analysis and Applications*, *3*(2), 358–360. https://doi.org/10.1016/0022-247X(61)90062-2

Bentley, A., Development, C., Prism, B., Bommel, E. Van, & Musser, E. (n.d.). Automation at scale is driving transformative change across insurance, 1–5.

Berkhin, P. (2006). Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data: Recent Advances in Clustering*, 25–71. https://doi.org/10.1007/3-540-28349-8_2

Bivand, R., Rundel, C., Pebesma, E., & Hufthammer, K. O. (2011). rgeos: Interface to Geometry Engine–Open Source (GEOS). *R Package Version 0.1-8*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Interface+to+Geometry+Engine+-+Open+Source+(GEOS)#0

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. *Use R* (Vol. 1). https://doi.org/10.1007/978-0-387-78171-6

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. *The Wadsworth statisticsprobability series* (Vol. 19). https://doi.org/10.1371/journal.pone.0015807

Brinkmann, P. (2016). Usage-Based Insurance : A European Case Study using Machine Learning.

Brondoni, S. M. (2002). Global Markets and Market-Space Competition. *Symphonya. Emerging Issues in Management*, (1), 28–42. https://doi.org/10.4468/2002.1.03brondoni

Brondoni, S. M. (2006). Corporate Communication and Global Markets. *Symphonya. Emerging Issues in Management*, (2), 9–37. https://doi.org/10.4468/2006.2.02brondoni

Brondoni, S. M. (2008). Ouverture de 'Market-Driven Management and Global Markets - 1'. *Symphonya. Emerging Issues in Management*, *0*(1), 1–13. https://doi.org/10.4468/2008.1.01ouverture

Brondoni, S. M. (2015). Global Networks, Outside-In Capabilities and Smart Innovation. *Symphonya. Emerging Issues in Management*, (1), 6–21. Retrieved from http://search.proquest.com/docview/1679397674?accountid=8144%5Cnhttp://sfx.aub.aau.dk/sfxaub?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ:pqrl&atitle=Glo

bal+Networks,+Outside-In+Capabilities+and+Smart+Innovatio

Cairo, A. (2012). Infographics and Visualization and exploration. *The Functional Art*, 15–25. Retrieved from http://www.thefunctionalart.com/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM -Cross-Industry Standard Process for Data Mining- 1.0 Step-by-step data mining guide. CRISP-DM Consortium*. https://doi.org/10.1109/ICETET.2008.239

Dang, J. (2017). Unveiling the full potential of telematics (p. 32). Swiss Re. Retrieved from http://media.swissre.com/documents/unveiling_the_full_potential_of_telematics_italy_case_study.pdf

Estivill-Castro, V. (2002). Why so many clustering algorithms - A Position Paper. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75. https://doi.org/10.1145/568574.568575

Everitt, B. (1980). Cluster analysis. *Quality and Quantity*, *14*(1), 75–100. https://doi.org/10.1007/BF00154794

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. *Quality and Quantity* (Vol. 14). https://doi.org/10.1007/BF00154794

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Focarelli, D. (2016). How the Connected World is Changing the Insurance Business : Lessons From Italy, (October).

Gorunescu, F. (2011). Data mining: Concepts, models and techniques. *Intelligent Systems Reference Library*, *12*. https://doi.org/10.1007/978-3-642-19721-5

Groupama Assicurazioni. (2016). *Groupama Assicurazioni Annual Report*.

Han, J., & Kamber, M. (2000). Data Mining: Concepts and Techniques. *Data Mining:Concepts and Techniques*, 3–26. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. *Elements*, *1*, 337–387. https://doi.org/10.1007/b94608

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. In *Procedia Engineering* (Vol. 100, pp. 816–825). https://doi.org/10.1016/j.proeng.2015.01.436

Italian AXA Paper. (2016). *Le sfide dei dati*. Retrieved from https://corporate.axa.it/documents/14601/108590/ITALIAN+AXA+PAPER+N.+8/fbf20c01-02f2-439d-8451-85aefb52697d

IVASS. (2008). Codice delle assicurazioni private. IVASS.

IVASS. (2017). *Annual Report*.

James, G., Witen, D., Hastie, T., & Tibshirani, R. (2007). *An Introduction to Statistical Learning with*

Applications in R. *Performance Evaluation* (Vol. 64). https://doi.org/10.1016/j.peva.2007.06.006

Jebara, T., & Jaakkola, T. (2000). Feature selection and dualities in maximum entropy discrimination. *Uncertainity In Artificial Intellegence*, 291–300. Retrieved from http://dl.acm.org/citation.cfm?id=2073981

Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Knaflic, C. (2015). *Storytelling with Data: A data visualization guide for business professionals*. (Wiley, Ed.).

Lovelace, R., & Cheshire, J. (2014). Introduction to visualising spatial data in R. *National Centre for Research Methods Working Paper 08/14*, *14*(3), 1–30. https://doi.org/10.5281/zenodo.889551

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management* (pp. 2–9). https://doi.org/10.1109/SSDM.1996.505910

Marabelli, M., & Newell, S. (2017). The Light and Dark Side of the Black Box: Sensor- Based Technology in the Automotive Industry. *Communications of the Association for Information Systems*, *40*(January). Retrieved from http://aisel.aisnet.org/cais/vol40/iss1/16

Marcucci, M., & Sharma, S. (1997). Applied Multivariate Techniques. *Technometrics*, *39*(1), 101. https://doi.org/10.2307/1270777

Mitchell, T. M. (1997). *Machine Learning*. *Annual Review Of Computer Science*. https://doi.org/10.1145/242224.242229

Pebesma, E. (2012). spacetime : Spatio-Temporal Data in R. *Journal of Statistical Software*, *51*(7), 1–30. https://doi.org/10.1359/JBMR.0301229

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R. J., Sumner, M., … Brien, J. (2016). Package 'sp'. *R*. https://doi.org/10.1016/j.jhydrol.2011.07.022.

Pebesma, E. J., & Bivand, R. S. (2005). Classes and Methods for Spatial Data in R. *R News*, *5*(2), 9–13. Retrieved from http://cran.r-project.org/doc/Rnews/

Porter, M. (1979). How Competitive Forces Shape Strategy. *Harvard Business Review*, *57*(2), 137–145. https://doi.org/10.1097/00006534-199804050-00042

Porter, M. E. (1985). Competitive Advantage. *Strategic Management*. https://doi.org/10.1108/eb054287

Porter, M. E., & Heppelmann, J. E. (2014). How Smart, Connected Product Are Transforming Competition. *Harvard Business Review*, (November), 64–89. https://doi.org/10.1017/CBO9781107415324.004

Porter, M. E., & Heppelmann, J. E. (2015). How smart, connected products are transforming companies. *Harvard Business Review*. https://doi.org/10.1017/CBO9781107415324.004

Provost, F., & Fawcett, T. (2013). Data Science for Business. *Book*. https://doi.org/10.1007/s13398-014-0173-7.2

Sharma, S. (1996). Applied Multivariate Techniques. *Technometrics*, *39*(1), 509.

https://doi.org/10.2307/1270777

Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., & Herdlein, S. a. (2000). The CRISP-DM model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22. Retrieved from www.spss.com%5Cnwww.dw-institute.com

Timm, N. H. (n.d.). *Applied Multivariate Analysis*.

Tufte, E. R. (1990). Data-Ink Maximization and Graphical Design. *Oikos*, *58*(2), 130–144. https://doi.org/10.2307/3545420

Tufte, E. R. (2001). The Visual Display of Quantitative Information. *Technometrics*. https://doi.org/10.1198/tech.2002.s78

Turban, E., Sharda, R., & Aronson, J. (2008). Business intelligence: a managerial approach. *Tamu-Commerce.Edu*, 1–30. https://doi.org/10.1109/HICSS.2012.138

Unwin, A. (2012). An Introduction to Applied Multivariate Analysis with R by Brian Everitt and Torsten Hothorn. *International Statistical Review*, *80*(2), 331–332. Retrieved from http://10.0.4.87/j.1751-5823.2012.00187_11.x%0Ahttp://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=78190220&site=ehost-live&scope=site

Wuthrich, M. V., & Buser, C. (2017). Data Analytics for Non-Life Insurance Pricing. *Swiss Finance Institute Research Paper*, 16–68. Retrieved from https://econpapers.repec.org/RePEc:chf:rpseri:rp1668

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/TNN.2005.845141

Zheng, Y. (2015). Trajectory Data Mining: An Overview. *ACM Trans. On Intelligent Systems and Technology*, *6*(3), 1–41. https://doi.org/10.1145/2743025

APPENDIX

Appendix 1

This appendix contains descriptive statistics computed over initial variables of the raw dataset. Variable "timestamp" is excluded from the following tables because it requires variables transformations to be applied before exploring them.

In Table 0.1 is presented the descriptive statistics for categorical variables with granularity to level.

Name	Level	Description	Frequency	FreqPercent
gps_quality	0	No Signal	377	0.00%
gps_quality	1	Marginal Signal	1648225	11.93%
gps_quality	2	2D Signal	44079	0.32%
gps_quality	3	3D Fix	12119450	87.74%
session	0	Engine On	696953	5.05%
session	1	In Movement	12418225	89.91%
session	2	Engine Off	696953	5.05%

Table 0.1 – Descriptive statistics for categorical variables

In Table 0.2 is presented the descriptive statistics for numerical variables.

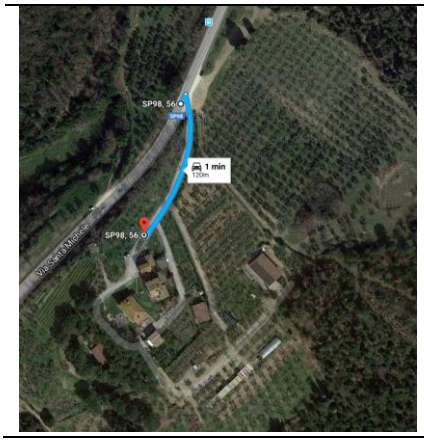
Variable Name	Minimum	Maximum	Mean	Std_Deviation	Skewness	Kurtosis
latitude	0	55.55	41.13	2.46	0.20	2.39
longitude	-9.03	28.93	13.54	2.59	-0.62	5.29
speed	0	288.07	33.32	30.34	1.20	4.47
heading	0	360.00	163.14	111.04	0.07	1.76
gps_quality	0	3.00	2.76	0.65	-2.33	6.43
distance_elapsed	0	65535.00	368.63	750.50	2.82	76.11
time_elapsed	0	99999999.00	526.22	100583.61	743.30	647433.46

Table 0.2 – Descriptive statistics for numerical variables

Appendix 2

This appendix contains the legend to interpret variable “road_type”:

Example	Label	Italian classification	Description
	**motorway**	Motorway	A restricted access major divided highway, normally with 2 or more running lanes plus emergency hard shoulder. Includes metropolitan’s ring roads like Tangenziali Esterne of Milan Great Ring Junction of Roma.
	**trunk**	Main Non-urban road	The most important roads in a country's system that aren't motorways. Need not necessarily be a divided highway but must be regulated with slip roads and without any crossing at grade.
	**primary**	State road or Regional road	The next most important roads in a country’s system. Link larger towns, normally two-way directions and can be not-separated by safety barriers.
	**secondary**	Secondary Non-urban road	The next most important roads in a country’s system. They are not main arterial routes but belongs to national road network because connects towns.
	**tertiary**	Local road	The next most important roads in a country’s system. In the national road network, they link smaller towns and villages. However, in OpenStreetMap “tertiary roads” can also connect minor roads to main roads.. Outside inhabited towns and villages, these roads are scares to moderate trafficked.

	**living_ streets**	Neighborhood streets	Roads which serve as an access to housing, without function of connecting settlements. Often lined with housing.
	**unclassified**	Oaks road	Minor roads of a lower classification than “tertiary”, but which serve a purpose other than access to properties. Often link villages and hamlets.
	**service**	Service road	For access roads to, or within an industrial estate, camp site, business park, car park etc. Next to main roads, enabling stop and grouping access from lateral properties to the main road and vice-versa, as well as movement and manoeuvres of vehicles not admitted on such main road.

Appendix 3

This appendix contains the correlation matrix based on Spearman's Index which is computed on ranks and depicts monotonic relationships. It is worth remembering that Peason's Index is computed on true values and capture linear relationships.

To explore data is best to compare results of the two indices because the relation between Spearman and Pearson correlations can give some information.

	DISTANCE_ 30_50_ avg_speed	DISTANCE_ 50_90_ avg_speed	DISTANCE_ 90_130_ avg_speed	DISTANCE_ 130_000_ avg_speed	n_ GPS	n_ behaviour	over_ speed	living_ street	motorway	primary	trunk	unclassified	tertiary	secondary
DISTANCE_30_50_avg_speed														
DISTANCE_50_90_avg_speed	0.32													
DISTANCE_90_130_avg_speed	0.10	0.38												
DISTANCE_130_000_avg_speed	0.02	0.11	0.32											
n_GPS	0.63	0.69	0.40	0.15										
n_behaviour	0.33	0.34	0.20	0.09	0.49									
over_speed	0.29	0.65	0.44	0.21	0.54	0.28								
living_street	0.01	0.00	0.02	0.01	0.06	0.05	0.11							
motorway	0.09	0.25	0.41	0.23	0.27	0.12	0.24	0.03						
primary	0.23	0.33	0.17	0.06	0.38	0.17	0.17	-0.02	0.09					
trunk	0.14	0.39	0.48	0.17	0.37	0.16	0.33	0.03	0.15	0.11				
unclassified	0.07	0.02	-0.01	0.00	0.06	0.06	0.05	-0.24	-0.04	-0.09	-0.01			
tertiary	0.11	0.03	-0.02	-0.01	0.12	0.08	0.13	-0.19	-0.03	-0.12	-0.02	-0.24		
secondary	0.18	0.18	0.04	0.01	0.25	0.13	0.04	-0.15	0.01	-0.10	0.01	-0.20	-0.27	

Figure 0.1 – Correlation Matrix using Spearman's Index for variables selected for final clustering model

It is interesting the correlation between "n_GPS" and "DISTANCE_30_50_AVG_SPEED" increases from 0.18 to 0.63. However, this is the only significant difference observed.

Appendix 4

For completeness of the thesis, it was computed an internal index for validating the number of cluster chosen. As previously mentioned, there are plenty of statistical indices but most of them are not easily interpretable or requires computing a distance matrix.

It was decided to use the R-Squared computed as the ratio between cluster sum of squares and total within-cluster sum of squares. This is a measure of the extent to which groups are different from each other. The value ranges from 0 to 1, with 0 indicating no difference among groups.

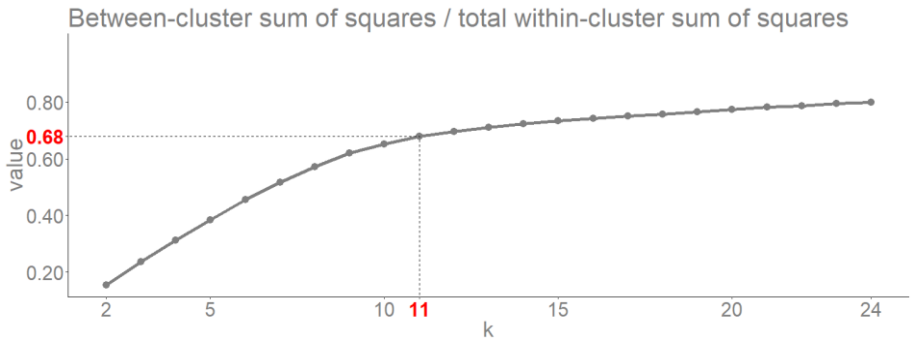


Figure 0.2 – Scree plot of R-Squared not presenting clear elbow

In Figure 0.2 there is no clear Elbow supplying statistical evidence for the choice of cluster 11.

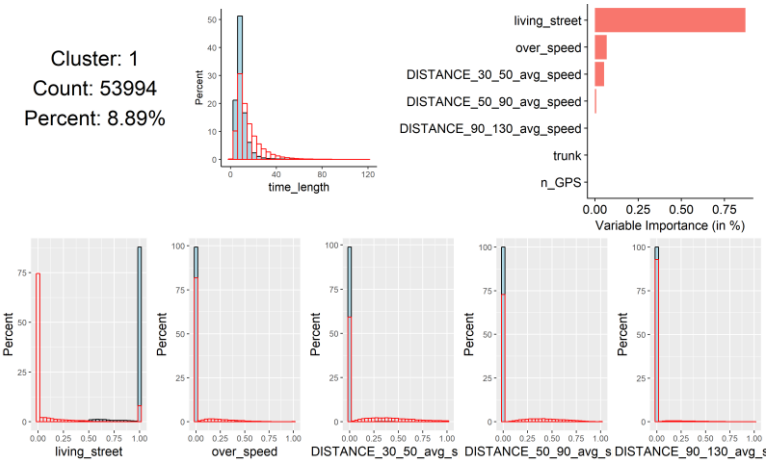
Appendix 5

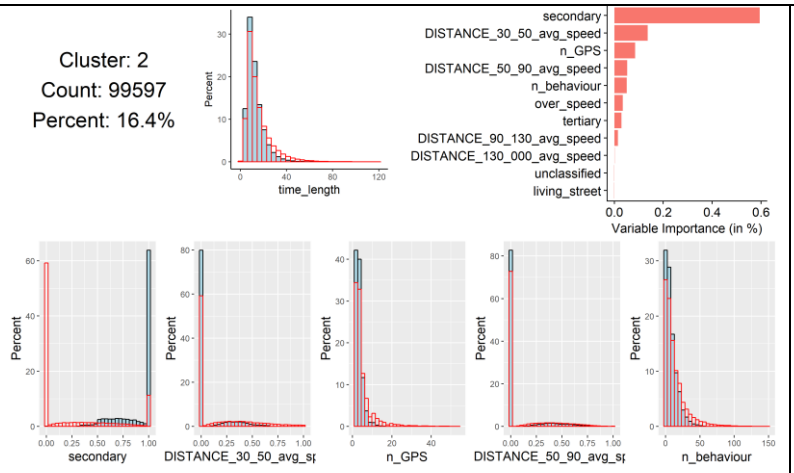
The following visualizations represent the output of a clusters profiling algorithm written for this project and inspired on the effective *Segment Profile* node of SAS Enterprise Miner.

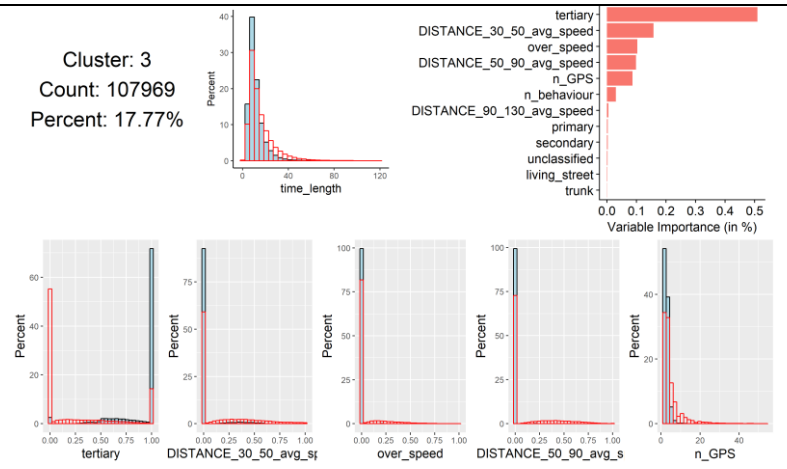
Variable "time_length" was not included in clustering because of highly correlated, however, it was decided to include it in order to supply additional information about the trip.

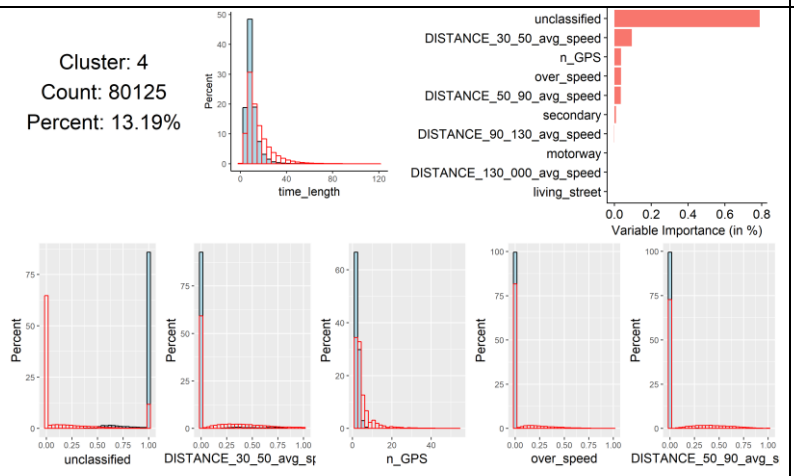
The horizontal bar chart on the top right-hand corner represent the variable importance of each variable computed by learning a binary classification tree to predict individuals of the cluster.

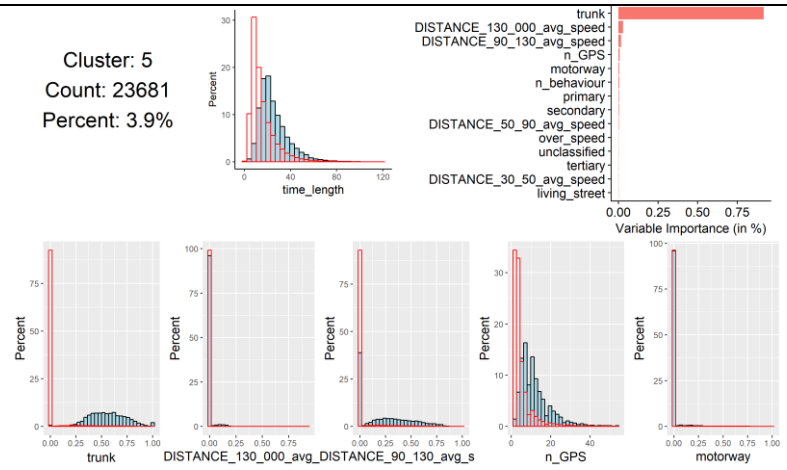
Each histogram on the bottom compare the distribution of the individuals in the cluster – blue colour – to the distribution of the overall population – red colour.



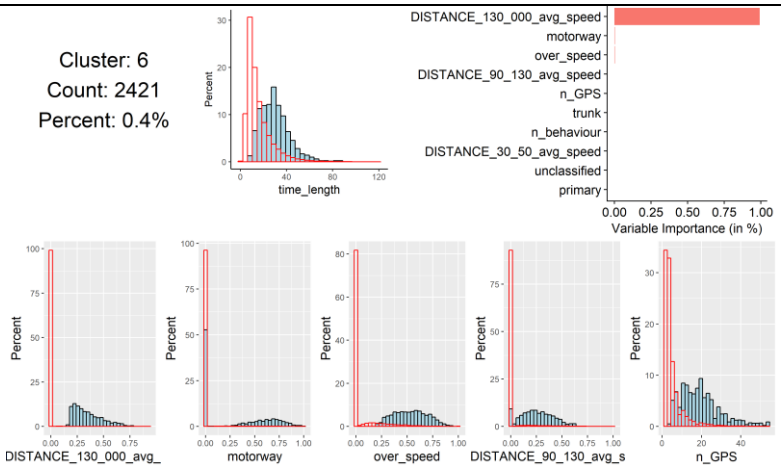




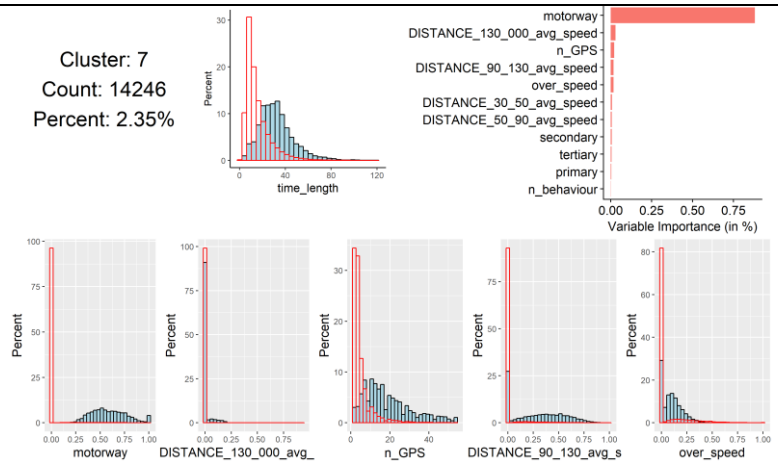




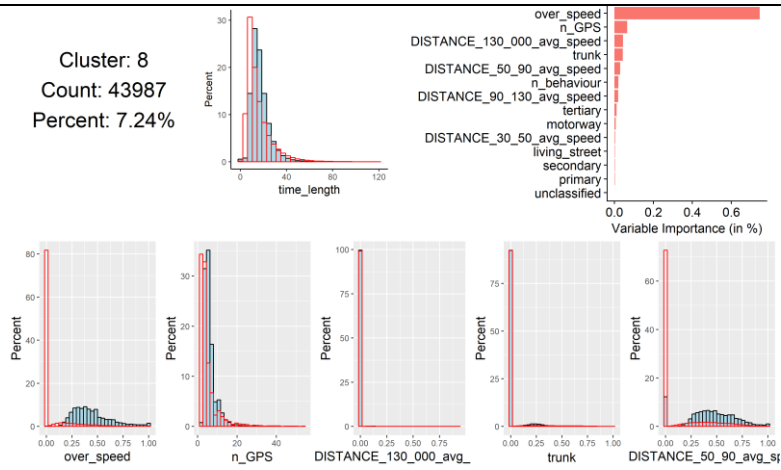
Cluster: 6
Count: 2421
Percent: 0.4%



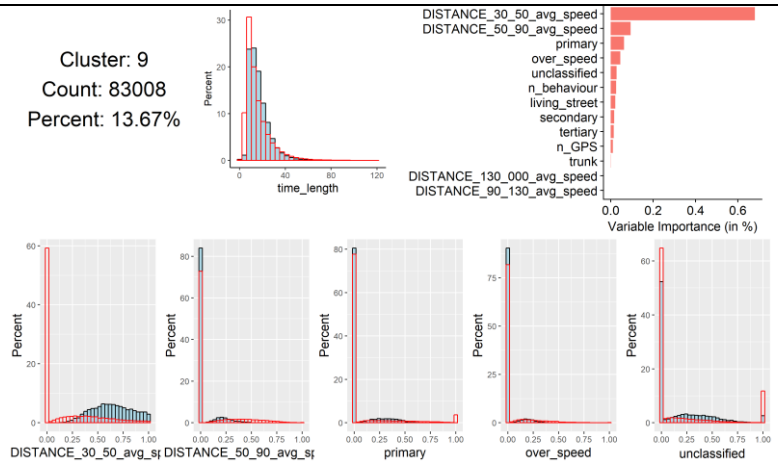
Cluster: 7
Count: 14246
Percent: 2.35%



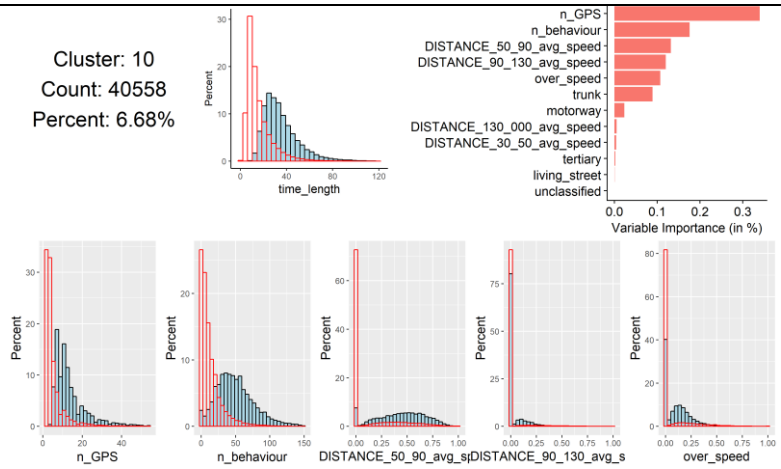
Cluster: 8
Count: 43987
Percent: 7.24%



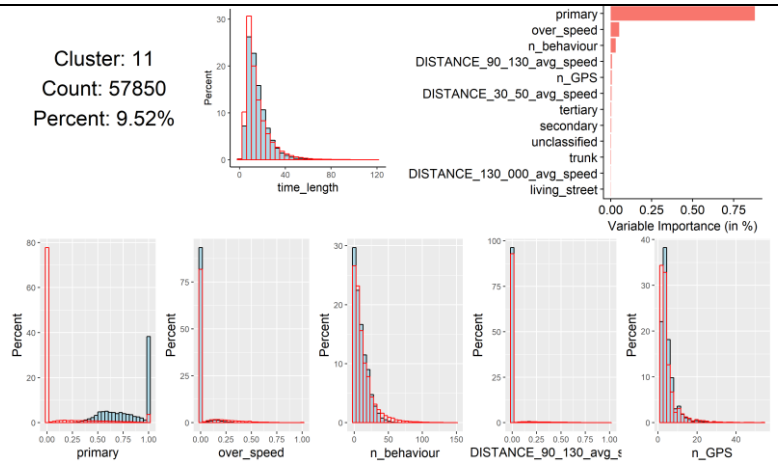
Cluster: 9
Count: 83008
Percent: 13.67%



Cluster: 10
Count: 40558
Percent: 6.68%



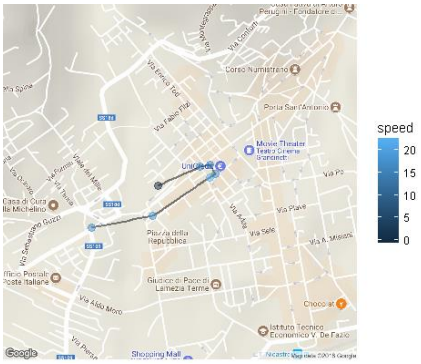
Cluster: 11
Count: 57850
Percent: 9.52%



Appendix 6

This Appendix visualizes trajectories of the most representative trip for each of the 11 clusters.

① Ordinary, very short trips exclusively on “living streets”

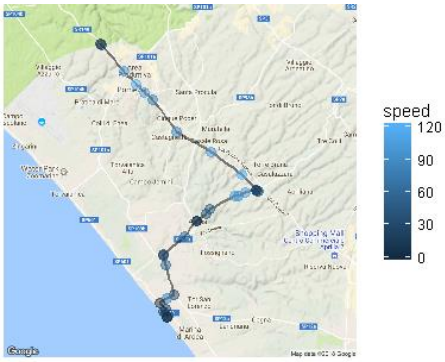


② Ordinary, short trips exclusively on "secondary" roads

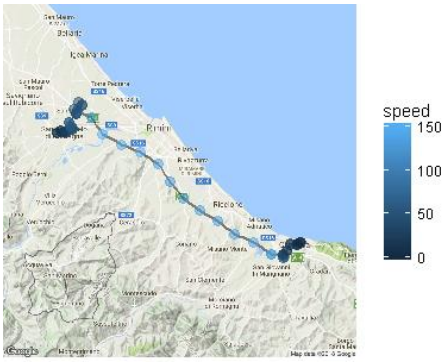
③ Ordinary, very short trips exclusively on “tertiary” roads

④ Ordinary, very short trips exclusively on “unclassified” roads

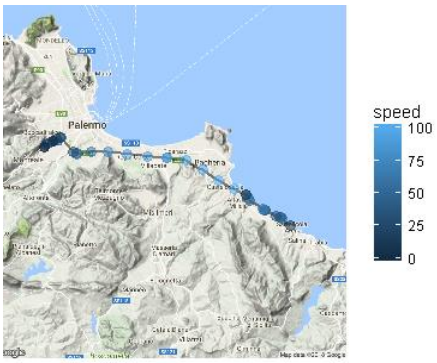
⑤ Long, fast trips on “trunk” roads



⑥ Extremely fast, long trips on “motorway” roads



⑦ Long, fast trips on “motorway” roads



⑧ Over speeding trips on “tertiary”, “secondary” and “unclassified” roads

⑨ Ordinary trips on “tertiary”, “secondary” and “unclassified” roads

⑩ Slightly over speeding trips with frequent behavioural events on “primary”, “secondary” and “tertiary” roads

⑪ Ordinary trips exclusively on “primary” roads