Ref. Ares(2023)6005345 - 04/09/2023

D1.3 Methods to extract statistically significant relationships between human/external factors and driver behavioural mechanisms, in uncritical and critical situations

D 1.3 | 19.04.2023

Deliverable No. 1.3 Data Mining Methods

Project Acronym	Grant Agreement #	Project Title	Deliverable Reference #	Deliverable Title
I4Driving	101076165	Integrated 4D Driver Modelling under Uncertainty	1.3	Methods to extract statistically significant relationships between human/external factors and driver behavioural mechanisms, in uncritical and critical situations

AUTHORS

Roberta Siciliano	University of Naples Federico II
Antonio D'Ambrosio	University of Naples Federico II
Michele Staiano	University of Naples Federico II
Andrea Saltelli	Institute for Cognitive Sciences and Technologies (Cnr-ISTC)
Giulia Vannucci	Institute for Cognitive Sciences and Technologies (Cnr-ISTC)
Monica Di Fiore	Institute for Cognitive Sciences and Technologies (Cnr-ISTC)

DISSEMINATION LEVEL

X	P	PUBLIC
	C	CONFIDENTIAL

Funded by the European Union

This project has received funding from the European Union's Horizon Europe programme, under grant agreement No 101076165.

Disclaimer

Funded by the European Union. Views and opinions expressed in this publication are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Version History

Revision	Date	Authors	Organisaton	Description
V0.1	21.03.2023	Vincenzo Punzo	Unina	First draft
V0.2	23.03.2023	Vincenzo Punzo	Unina	Second draft
V0.3	30.03.2023	Vincenzo Punzo	Unina	Third draft
V0.4	19.04.2023	Lucia Schlemmer	Panteia	Final draft
V1.0	10.05.2023	Maria Rodrigues	Panteia	Final version

Executive Summary

The major scientific challenge of the **i4Driving Project** is developing a human driver model that captures the relevant behavioural mechanisms for safety assessment. The project has a specific goal to identify causal relationships among external, human factors and safety-critical driver behaviours at the level of specific driving situations.

The key question is how factors like *"gender, cultural and ethnic background, ageing, impairments, driving experience and route familiarity, mental workload or fatigue, weather and lighting conditions"* are statistically correlated with safety-critical driving behaviours. This deliverable defines the data mining framework to:

- Explain the cognitive/perceptive process of human driving;
- Predict the causes that lead to safety-critical driving behaviours; and
- Explain the causes exogenous to human behaviour that are related to road accidents and the same driving behaviours that are dangerous for safety.

Additionally, it introduces to global sensitivity analysis methods that are used for selecting features, following recent developments on the use of sensitivity analysis of and for data mining (Tunkiel, Sui, and Wiktorski 2020; Antoniadis, Lambert-Lacroix, and Poggi 2021), also including the use of the concept of mean dimension (Hoyt and Owen 2021).

Data Mining and Machine Learning can be useful considered for Driving Behavioural Analysis. The input is to consider external factors, human factors and safety-critical driver behaviours. The output is to identify a set of causal relationships, observable features in the data. Data mining methods aim to detect discriminant key human factors and safety-critical driver behaviours under specific driving conditions. Machine Learning Modelling can be applied to identify significant causal relationships between external and human factors, and safety-critical driver behaviours.

The formulation and structuring of hypotheses is the basis in the machine learning modelling, as well as in the experimental design. Selecting a comprehensive set of use cases and scenarios requires both Data Scientists and Domain Experts.

Statistical Learning and Data Analysis using Data Mining methods have been performed in a case study. The databases on mobility deriving from the **Strategic Research Program (SHRP2)** on Naturalistic Driving Study collected by University of Virginia have been considered (https://insight.shrp2nds.us). The aim is to address the driver performance and behaviour in traffic safety.

Contents

Executive Summary 3
1. Introduction 6
1.1 Statistical Methodological Vision 6
1.2 Data Mining 7
1.2.1 Data Mining and Knowledge Discovery 8
1.2.2 CRoss-Industry Standard Process for Data Mining (CRISP-DM) 9
1.2.3 Data Mining in Total Quality Management 10
1.3 Data Sources and Data Types 14
1.4 Application Scenarios and Methodological Building-Blocks 16
2. Exploratory Data Analysis 19
2.1 Association Rules Mining 19
2.2 Dimensionality Data Reduction 19
2.3 Clustering Methods 20
3. Modelling and Prediction 22
3.1 Supervised Learning Methods 22
3.2 The bias-variance dilemma 22
3.3 Interpretable versus Black Box Machine Learning 23
4. Sensitivity Analysis 26
4.1 Uncertainty versus Sensitivity Analysis 26
4.2 Variance-based sensitivity analysis 26
4.3 Sensitivity analysis in practice 27
4.4 Integrate GSA with machine learning algorithms 29
4.5 Sensitivity auditing 29
5. Case Study on SHRP2 Databases 30
5.1 Exploratory Data Analysis of the "Event" dataset 30
5.1.1 Data Preparation 30
5.1.2 Multiple Correspondence Analysis 33
5.2 Modelling and Prediction 38
5.2.1 Classification Trees 38
5.2.2 Random Forests 41
6. Next steps 41
References 43

List of Figures

Figure 1. Data Mining and Knowledge Discovery (Nisbet, Elder & Miner, 2009). 8

Figure 2. CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Martínez-Plumed et al., 2019). 9

Figure 3. Knowledge Discovery Pyramid (Siciliano and D'Ambrosio, 2012). 11

Figure 4. Statistical Learning Process in Total Quality Management (Siciliano and D'Ambrosio, 2012). 13

Figure 5. Trade-off Bias-Variance. 23

Figure 6. Trade-off Interpretability-Flexibility (James, Witten, Hastie & Tibshirani, 2009). 24

Figure 7. Scatterplots with moving averages (red). The straight line is the standardized regression coefficient of y on x_i, the discontinuous line is the moving average E_(x_(~i)) (y | x_i) (Saltelli, 2008). 27

Figure 8. Factorial Representation of Multiple Correspondence Analysis (first and second dimensions). 36

Figure 9. Factorial Representation of Multiple Correspondence Analysis (third and fourth dimensions). 37

Figure 10. Factorial Representation of Multiple Correspondence Analysis (fifth and sixth dimensions). 38

Figure 11. Classification Tree (Target Variable: Event Severity). 40

Figure 12. Random Forests Predictor Importance (Target Variable: Event Severity). 41

List of Tables

Table 1. Frequency Distribution of the Event Severity. 30

Table 2. Statistical Descriptive Summary of the Event Start. 30

Table 3. List of Categorical Variables with their number of categories. 31

Table 4. List of categories to be merged into "other". 32

Table 5. Cumulative Inertia of Factors of Multiple Correspondence Analysis. 34

Table 6. Benzecrì Correction of Cumulative Inertia of Factors of Multiple Correspondence Analysis... 35

Table 7. Contribution of the most important categories to the factorial representation (first and second dimensions). 35

Table 8. Contribution of the most important categories to the factorial representation (third and fourth dimensions). 36

Table 9. Contribution of the most important categories to the factorial representation (fifth and sixth dimensions). 37

Table 1. Frequency Distribution of the Event Severity. 31

Table 2. Statistical Descriptive Summary of the Event Start. 31

Table 3. List of Categorical Variables with their number of categories. 32

Table 4. List of categories to be merged into "other". 33

Table 5. Cumulative Inertia of Factors of Multiple Correspondence Analysis. 35

Table 6. Benzecrì Correction of Cumulative Inertia of Factors of Multiple Correspondence Analysis.. 36

Table 7. Contribution of the most important categories to the factorial representation (first and second dimensions). 36

Table 8. Contribution of the most important categories to the factorial representation (third and fourth dimensions). 37

Table 9. Contribution of the most important categories to the factorial representation (fifth and sixth dimensions). 38

1. Introduction

This report details the selection and use of specific machine learning techniques for identifying the statistically significant relationships between factors and behaviours. It presents the outcome of task 1.2. Mining data for behavioral modelling.

1.1 Statistical Methodological Vision

Road mobility is constantly evolving, posing new challenges for science, industry, public bodies and policymakers. New (automated) and classic (human-driven) vehicles will soon find themselves coexisting which will pose various problems, namely around road safety. An accurate assessment of the causes of accidents and the level of road safety is thus crucial for safe and efficient mobility.

The application context of road accidents and their prevention is investigated in the **i4Driving Project**. One fundamental task is to frame the statistical methodologies and the process of data analysis that can be fruitfully used to extract significant relationships between human/external factors and driver behavioural mechanisms, in uncritical and critical situations.

The methodological framework is based on **Data Mining**: its definition and its impact in the digital era will be discussed in Section 1.2. Data Mining (Hand, 1998) means *"digging in the mine of data"* when they are huge, high dimensional and structured in complex way. It requires an actionable strategy defining the process to collect, organise and prepare data for model building and derive useful outcomes and value for the stakeholders and beneficiaries of the data analysis.

One approach is offered by the view of Data Mining as the core of the **Knowledge Discovery Process (KDD)** (Nisbet, et al., 2009) as discussed in Section 1.2.1. Another strategy is provided by the **Cross Industry Standard Process for Data Mining (CRISP-DM)** (Martínez-Plumed et al., 2019) as described in Section 1.2.2. In both strategies, one step considers the model building where exploratory as well as machine learning methods can be fruitfully considered. A third actionable strategy of Data Mining allows to assure total quality management in all steps of DM process (Siciliano and D'Ambrosio, 2012) as discussed in Section 1.2.3.

Fundamental is to identify two key roles with their specific tasks:

- The ***Domain Expert,*** to formulate the challenging questions, the desired outcome, how to understand the facts of interest and the associated data.
- The ***Statistician or Data Scientist,*** to formulate the strategy of data analysis, the way to process data and how to interpret the results.

The definition and use of statistical methodologies, based on Data Mining and Machine Learning (often labelled as data-driven modelling), for the analysis of mobility problems cannot ignore the availability of large and quality data sources certified.

The **i4Driving Project** deals with a variety of **data sources and data types,** discussed in Section 1.3. Specifically, the diversity in data comes from the variety of sensors and connected vehicles, and even cyclists and pedestrians who can contribute to generating localised data streams due to their smartphones, as well as data accumulated via road surveillance systems or collected for **naturalistic** and **simulator studies**.

This deliverable is structured as follows:
The **methodological building blocks of Data Mining** with the **application scenarios for i4Driving Project** will be discussed in Section 1.2. **Exploratory Data Analysis** will be discussed in Section 2.
The Statistical Learning paradigm (Vapnik, 1995, 1998) and its declination in **machine learning models for inference and prediction** (Hastie et al., 2009; James et al. al., 2021) will be discussed in Section 3 considering the condition of application, the model assessment, the model selection criteria.
Validation of data mining and machine learning models can be investigated by **Sensitivity Analysis (SA)** as discussed in Section 4.
A case study will be discussed in Section 5. The database is derived from the **Strategic Research Program (SHRP2)** on Naturalistic Driving Study collected by University of Virginia to address the driver performance and behaviour in traffic safety. This case study allows to show the potentialities of Data Mining and Machine Learning methods in **i4Driving Project** domain applications, specifically ***to extract statistically significant relationships between human/external factors and driver behavioural mechanisms, in uncritical and critical situations.***

1.2 Data Mining

The British statistician David Hand was among the main promoters of **Data Mining -** a new frontier of statistical methodology – which is defined as *"the process through which the use of non-trivial models aims to identify relationships between non-trivial, hidden, useful and usable data"* (Hand, 1998).

What was new about Data Mining?
In the primary data analysis, the data are collected with a particular question or set of questions in mind. This primary statistical analysis works with datasets that are small and clean, very often single, static datasets and sampled in an *independent, identically distributed* manner, collected to answer the problem being addressed, and which are solely numeric.

However, not all datasets fit this description. Often the datasets can be:

- Very large in size;
- Contaminated;
- Have features of non-stationarity, selection bias and dependent observations;
- The interest is in finding interesting patterns rather than identifying a stochastic theoretical model generator of all data;
- Data are not only numeric, but also categorical and mixed;
- There can be spurious relationships; and
- Due to the complexity of the data structure, there is a need for Automated Data Analysis.

Computer technology and electronic data acquisition yields to the growth of huge, more complex databases. These are viewed as a resource in different domains of application. With more complex data, data mining is necessary. Data mining can be defined as follows:

"*Data Mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners*" (Hand, 1998).

"*Data Mining is the study of collecting, cleaning, processing, analysing, and gaining useful insights from data. It can be organised into a process used to extract usable information from a larger set of any raw*

data. It implies analysing data patterns in large batches of data using one or more software" (Aggarwal, 2015).

Data Mining is thus a strategy of analysis of complex data structures aiming to discover significant relationships, to classify similar data and to visualise key points of interest or value for the owners of the data and the stakeholders.

Pioneer contributions to Data Mining were Tukey's Exploratory Data Analysis (EDA) (Tukey, 1977) and Benzecrì's school of the *Analyse des données* in France (Lebart, Morineau, Warwick, 1987). The starting point was the implausibility of probabilistic assumptions of multivariate data analysis and the need of a data-driven approach. The aim was to optimise the graphical visualisation of multidimensional data and to provide the complementary use of more methods in a strategy of analysis, nowadays known as Data Mining. As an example, one best practice is a two-stage strategy. In the first stage, a factorial method is applied, such as **Principal Component Analysis,** to reduce the dimensionality of the dataset by eliminating the redundancy of information resulting from highly correlated variables and by replacing them with a smaller number of new latent variables that are not correlated with each other and linearly linked to the starting variables. In the second stage, a **Cluster Analysis** on the latent variables is applied, to detect in an optimal way clusters of similar data. Visualisation and additional aids for the interpretation of the results provide useful information for the beneficiary of the data analysis.

1.2.1 Data Mining and Knowledge Discovery

Data Mining is a phase of the broader process called **Knowledge Discovery from Databases (KDD)** with the aim of extracting useful information, taking full advantage of the information deriving from ever-increasing amounts of data available in the digital epic of bits.

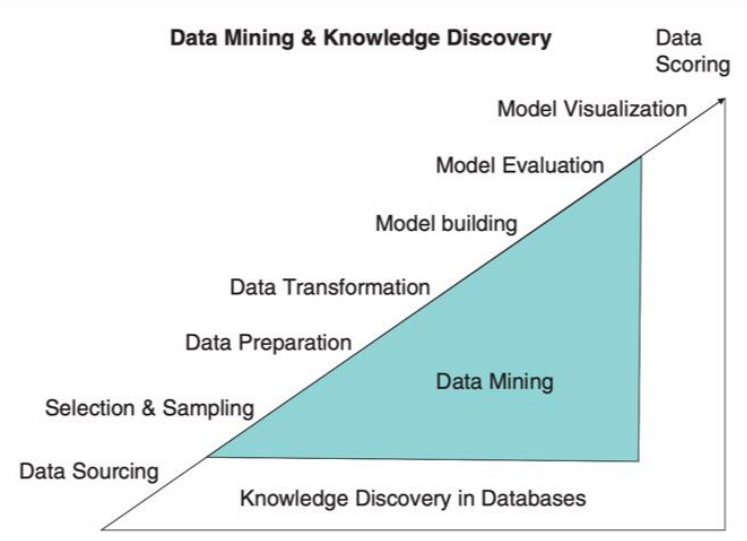

Figure 1. Data Mining and Knowledge Discovery (Nisbet, Elder & Miner, 2009).

There are two fundamental steps as part of the Data Mining and Knowledge Discovery process:

- ***Selection and Sampling***, which refers respectively to features selection and sampling data points; and

- ***Data Preparation***, to convert the data into a form suitable for further processing, namely that it is directly ready for machine learning models.

1.2.2 CRoss-Industry Standard Process for Data Mining (CRISP-DM)

Recently, a **CRoss-Industry Standard Process for Data Mining (CRISP-DM)** (Martínez-Plumed et al., 2019) was proposed. This is an open standard process model that describes common approaches used by data mining experts. Significant efforts concern the **Data Collection** and **Data Preparation** phase.

CRISP-DM is a cyclical process (figure 2) that starts from the formulation of the real problem in a particular application context (*Domain Understanding*) and the understanding of the available data (*Data Understanding*). It then moves on to the phases of data pre-treatment (*Data Preparation*) and definition of statistical learning models (*Modelling*). It is completed with the phases of evaluation (*Evaluation*) and analysis of impact and sensitivity in the real context (*Deployment*). Within the phase of 'Modelling', various statistical learning models can be considered.

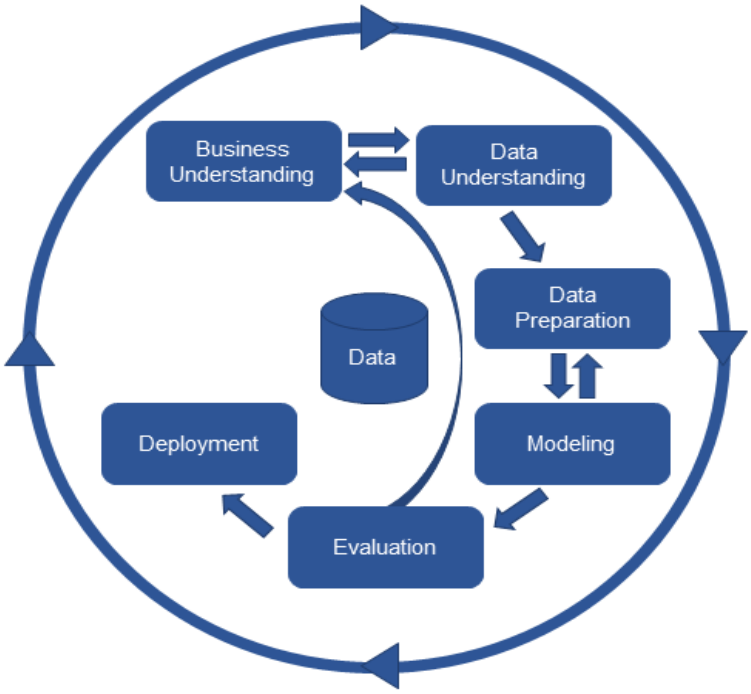

Figure 2. CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Martínez-Plumed et al., 2019).

Key points of consideration in CRISP-DM are:

- **Domain Understanding:** What are the challenging questions?
- **Data Understanding:** What data do we have or need?
- **Data Preparation:** How do we organize the data for modelling?
- **Modelling:** What modelling techniques should we apply?
- **Evaluation:** Which model best meets the business objectives?
- **Deployment:** How do stakeholders access the results?

1.2.3 Data Mining in Total Quality Management

There is a third strategy called **Statistical Learning and Information Management (SLIM)** to manage all steps of Data Mining, to assure quality. The pioneering contribution was introduced in the "Digital Accessible Statistical Information System for Monitoring Tourism" in Campania Region (Siciliano and D'Ambrosio, 2012). The SLIM strategy has been considered in the MAGIC European Project[1] to rationalise the data management and statistical data analysis process. SLIM was revealed powerful enough to satisfy various research needs in three different domain applications (Energy, Water, Food) and capable of dealing with a variety of data collection and data structures, reinforcing the team building made up of statisticians and domain experts. SLIM can be updated to the specific needs of the **i4Driving Project**.

Two conceptual models are considered:

- The ***Knowledge Discovery Pyramid* (*KDP*).**
- The ***Statistical Learning Process* (*SLP*).**

The ***Knowledge Discovery Pyramid* (*KDP*)** emphasises the key steps to provide added value in decision-making, starting from the challenging questions. It consists of the following stages (figure 3):

- **Real problem**: challenging questions submitted by the beneficiary.
- **Brainstorming**: real problem converted into statistical challenges.
- **Data sharing**: share all useful and available data.
- **Data accessibility/integration**: collect other data.
- **Filtering**: selection of features and units.
- **Research**: strategy of statistical learning and data analysis.
- **Processing**: run methods and algorithms.
- **Analysis**: output interpretation and validation.
- **Results exploitation**: prediction and decision-making.
- **Value**: outcome and significant actions to be implemented.

[1] **MAGIC** (H2020-EU.3.5.4 - G.A. 689669), Moving Towards Adaptive Governance in Complexity: Informing Nexus Security – 2016- 2020 - https://magic-nexus.eu.

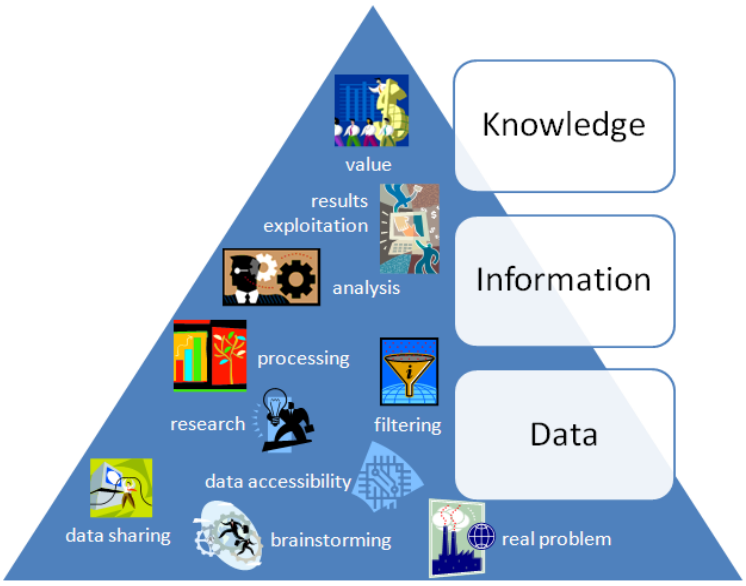

Figure 3. Knowledge Discovery Pyramid (Siciliano and D'Ambrosio, 2012).

The stages in the ***KDP*** rationale involve two key roles:

- The ***Domain Expert***, the beneficiary of the quantitative knowledge discovery, who formulates the challenging questions, the desired outcome, how to understand the facts of interest and the associated data.
- The ***Data Scientist*** or Statistician, who formulates the strategy of data analysis, the way to process data, how to interpret the results.

There are three levels in KDP as it moves *from Data to Information* and *from Information to Knowledge.*

1st level DATA includes the stages:

Real problem → Brainstorming → Data sharing → Data accessibility → Filtering

This involves both the Data Scientist and the Domain Expert to provide new entries for the real-world case study. It aims to overcome the potential clash between a 'question driven' (i.e., narrative/grammar based) and a 'data driven' (i.e., *a priori* framed) approach to ignite the Quantitative Storytelling.

The Domain Expert defines the *real problem and challenging questions* using any prior qualitative information and constraints, theory, and research hypotheses.

The Data Scientist acts in *brainstorming* with the Domain Expert to transform the real problem into one or more statistical challenges such to be faced quantitatively.

2nd level INFORMATION includes the stages:

Research → Processing → Analysis

The Data Scientist applies the *research* in statistical methodology for *processing* available data, sometimes requiring a strategy of analysis with the complementary use of standard methods, pre-processing of data, the use of nonstandard methods, new methodological development to match specific requests.

The *Analysis of Data* provides the final output, and the interpretative issues handled by the Data Scientist and the Domain expert together.

3rd level KNOWLEDGE includes the stages:

Results exploitation → Value

The Data Scientist and Domain Expert together provide the *results exploitation*, by outlining the answers to question marks, simplifying the output of the statistical analysis, summarising the results, formalising quantitative storytelling and the way to visualise and publish the results.

The result of the overall discovery process is *knowledge*, which can be fruitfully used to provide added value to the beneficiaries, which can be also measured in terms of impact or simply decision-making and accurate scenario prediction, suited for sensitivity analysis. The beneficiaries are the stakeholders, the policymakers and eventually the public.

One of the main results of brainstorming activity involving the Domain Expert and the Data Scientist is to understand which sort of data is necessary to face the real problem with statistical learning.

Three are the possible scenarios:

- Data are not available and need to be produced (i.e., simulator survey to design).
- Data are available (i.e., from databases, data warehousing, etc.) and need to be extracted.
- Data are available and complete, only need to be analysed.

Deming's cycle ***"Plan-Do-Check-Act"*** of Total Quality Management (Deming, 1950) introduced for industrial processes, successfully facilitates the managing and quality assurance of Statistical Learning Process.

The ***Statistical Learning Process*** **(*SLP*)** is structured into *three Deming cycles of Total Quality Management* to help in identifying the start-up of statistical analysis and which cycle of distinct activities needs to be followed up.

The **Statistical Learning Process** consists of three Deming's cycles (figure 4):

⇒ **Inner Cycle**: *Data Production*, to directly collect the dataset.

⇒ **Intermediate Cycle**: *Data Extraction*, to access available databases and filter units and variables of interest.

⇒ **Outer Cycle**: *Data Analysis*, to process data and find the results.

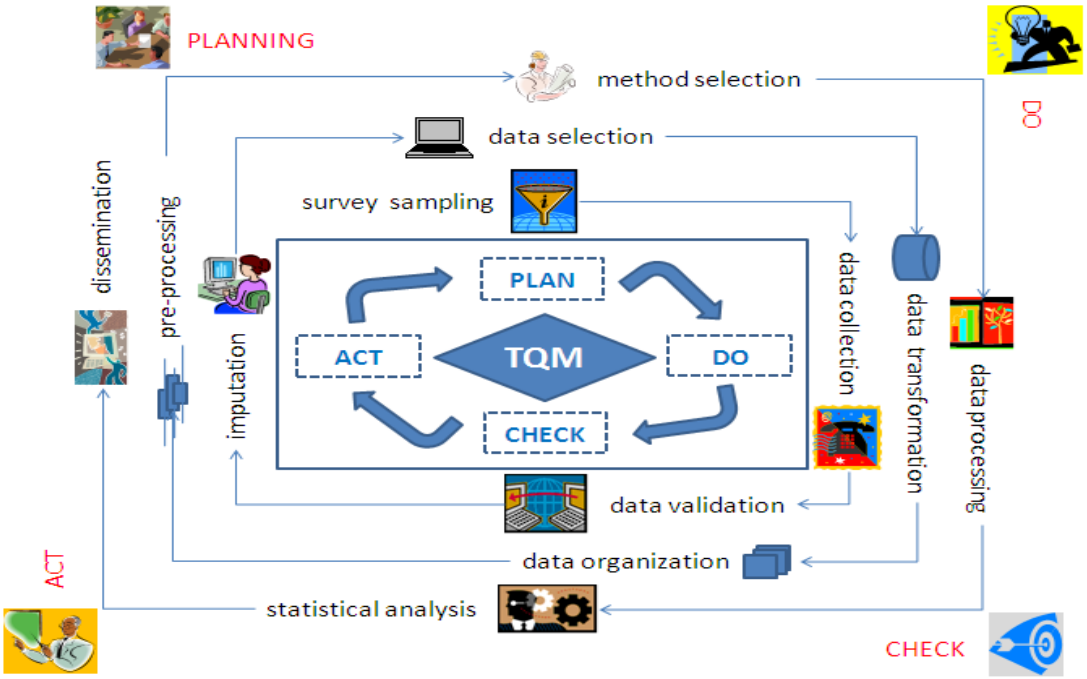

Figure 4. Statistical Learning Process in Total Quality Management (Siciliano and D'Ambrosio, 2012).

.

The **Data Production (inner) Cycle** is made up of the following steps:

- **Plan:** *Survey Sampling* for planning the way to collect data,
- **Do:** *Data Collection* for doing the operative part of field analysis,
- **Check:** *Data Validation* for assuring the quality of collected data,
- **Act:** *Data Imputation* for replacing missing data.

The **Data Extraction (intermediate) Cycle** is made up of the following steps:

- **Plan:** *Data Selection* for filtering units and variables of interest,
- **Do:** *Data Transformation* for recoding and normalization,
- **Check:** *Data Organization* for assuring the coherence of data of interest,
- **Act:** *Pre-Processing* for preliminary descriptive statistics.

The **Data Analysis (outer) Cycle** is made up of the following steps:

- **Plan:** *Method Selection* for choosing the statistical methodology,
- **Do:** *Data Processing* for providing the output of statistical reports,
- **Check:** *Statistical Analysis* for interpreting the output,
- **Act:** *Dissemination* for sharing the results.

It is iterative by nature so that any cycle can be repeated to improve and provide additional results.

As an example, to implement a survey study using the simulator, the start-up is the first cycle, whereas to analyse the data bases of the Naturalistic Driving Study already collected, the start-up is the second cycle. Once the data matrix is ready for the statistical data analysis, the start-up is the third cycle.

To implement all steps, there are some fundamental requirements:

- **Documentation Requirement**: Define the Input and Output documentation at any step.
- **Assignment rule**: Identify three key roles:
  - ✓ ***Data Sherpa for Data Production***
  - ✓ ***Data Harvester for Data Extraction***
  - ✓ ***Data Analyst for Data Analysis***
- **Assessment and Pedigree:** Certificate the quality of data management and data analysis, such to guarantee pre-producibility[2] and reproducibility.

The concept of reproducibility has been recognized as essential for the Post-Normal Science setting, and included in the ***NUSAP framework***, by Ravetz and Funtowics (1998): *"only when there is effective quality control of science for policy, through the management of uncertainties, will we be able to cope intelligently with the crises we face".* Saltelli (2020) in the fundamental contribution on ***Ethics of Quantification*** states: *"Any number that does not represent its context and purpose of production runs the risk of obfuscating as much as illuminating".*

NUSAP is a notational system for *the management and communication of uncertainty in science for policy*, based on five categories for characterizing any quantitative statement:

- **Numeral** will usually be an ordinary number.
- **Unit** refers to the units used in Numeral.
- **Spread** is an assessment of the error in the value of the Numeral.
- **Assessment** is a summary of salient qualitative judgements about the information – this can be of statistical nature (a significance level) or more general, i.e., involving terms such as 'conservative' or 'optimistic'.
- **Pedigree** is an evaluative description of the mode of production and of anticipated use of the information.

The pedigree can be expressed by means of a matrix; the columns represent the various phases of production or use of the information, and each column contains marks to rank the performance. The key goal of NUSAP method is ensuring transparency and communicability of quantitative information to promote trust.

1.3 Data Sources and Data Types

Challenging questions in the **i4Driving Domain** requiring Data Mining are around the data. **Data** results from the measurement of a *variable* (or feature) on a *unit* (or object, individual) under observation or experimentation at a given time and space.

The **size** of a dataset refers to the number of units, whereas the **dimensionality** of the data refers to the number of variables. Typically, **multidimensional data** are analysed.

[2] *"An experiment or analysis is preproducible if it has been described in adequate details for others to undertake it"* (Stark, 2018).

Based on the scale of measurement, we distinguish between **quantitative (or numerical) variables** and **qualitative (or categorical) variables**. A further distinction is made between **ordinal** and **nominal** variables if the categories of the qualitative variable can be ordered or not.

Coding is the recording of the measurement with a "value" so that the "value" can have its own value number, which can be processed in the case of quantitative variables, or, in the case of a qualitative variable, to indicate a certain category which is then suitable for the classification of data points.

Raw data is interpretable as a row vector or **record of values** corresponding to the measurements of all variables of the unit.

There is an information hierarchy in the scale of measurement of a variable, where the numerical variable scale has the highest information value over all the other types due to the data processing methods that can be applied.

The data processing of numerical variables exploits all information contents of the variable, not only their diversity or ordering, but also the intrinsic value of the coding operation. It is fundamental to understand the unit scale of the variable. Data transformation and normalisation operations might occur before data processing multidimensional data.

There is a drawback to such completeness of information. Outliers may occur, in which case anomaly detection procedures and intervention strategies may be necessary in data pre-processing.

Data can be extracted by various **data sources** and can be of **different types**, such as standard data, image, text, graph, data stream, preference rankings, symbolic data (interval data, histogram data, etc.), etc. Some examples for **i4Driving Project** are:

- **Surveys:** standard data collected directly and organised in a data matrix where the rows are the units and the columns are the variables.
- **Sensors, Internet of Things devices, tracking devices and other smart devices** collecting **data streams** (a sequence of digitally encoded coherent signals, often organised into a series of "packets").
- **Satellites** collect images and data using cameras.
- **Geographic maps** providing **spatial data.**

Non-standard data types can be transformed into standard data types in the data pre-processing step.

The fundamental distinction in Data Mining is between Non-Dependency-Oriented Data and Dependency-Oriented-Data.

Non-dependency-oriented data do not have any specified dependencies between either the data records or the variables.

- Binary data can be considered as a special case of either a categorical data or a numerical variable.
- Set-wise data is a set element indicator function to express the condition when its value is 1 and not 0.
- Text data is a string, a dependency-oriented data since it is a sequence of characters (or words) corresponding to the document, namely a vector-space representation where frequencies of the words (i.e., terms) are processed. This is typical in Natural Language Processing.

In **dependency-oriented data,** there might be implicit or explicit relationships between data records:

- Implicit dependencies means that dependencies are known to exist "typically" (i.e., sensor): If the temperature value recorded by a sensor at a particular time is significantly different from that recorded at the next time instant is extremely unusual and may be interesting for the data mining process.
- Explicit dependencies means that dependencies are explicitly specified, such as in graph or network data where edges are used to specify the relationships.

Time-series data contains values generated by continuously measurements over time (i.e, environmental sensor, speed sensor, etc.). This can be characterised by either:

- **Contextual variables** to define the context, based on which the implicit dependencies occur in the data:
  - in sensor data, the time stamp of the reading,
  - in spatial data, the location, and other characteristics of the reading.
- **Behavioural variables** to represent the values that are measured in a particular context (i.e, temperature): multiple sensors record readings at synchronised time stamps yields to a multidimensional time-series dataset.
  - Two sensors at a particular location monitor the temperature and pressure every second for a minute. This yields a bi-dimensional series of 60 points. The time stamps be replaced by index values from 1 through 60, especially when the time-stamp values are equally spaced apart. Time-series data are relatively common in many sensor applications and forecasting scenario applications.

Discrete Sequences and Strings are the categorical analog of time-series data.

Spatial data considers behavioural variables (i.e., sea-surface temperature) as well as contextual variables (i.e., spatial coordinates)**.**

Spatiotemporal data may consider either both spatial and temporal variables as contextual **(**i.e., variations in the speed-space is measured over time) or the temporal variable as contextual whereas the spatial variables as behavioural (i.e., **the trajectory analysis**).

1.4 Application Scenarios and Methodological Building-Blocks

In the framework of the ***i4Driving domain***, there is a surge of diverse data due to widespread sensors and connected vehicles, cyclists and pedestrians (that can contribute to generating localised data streams via their smartphones), as well as data accumulated via road surveillance systems or collected for naturalistic and simulator studies. The data sources referred to are twofold:

- Databases on mobility collected in naturalistic driving conditions; an example is represented by data derived from the **Strategic Research Program (SHRP2)** on the Naturalistic Driving Study, collected by University of Virginia, to address the driver performance and behaviour in traffic safety (a case study application will be shown in Section 5).
- Data collected by driving **simulation systems**.

The application context of road accidents and their prevention is investigated in the **i4Driving Project**. To date, there is no shared and uniformly adopted standard on the record route that defines *"the road accident"* as well as shared coding in the measurement of the variables of interest. At the same time, there is also an enormous difference in the quality of the data collected between the various

institutions and the official bodies in charge. The same applies to the data collected by driving simulators. The different systems are characterised by different methods of data collection, organisation of experiments and kinematic systems.

A specific research task in the i4Driving Project focuses on ***"Data Preparation"***. This phase means to fix up a set of pre-treatment activities of the data sources aimed at obtaining databases that:

- Harmonise the data from the various original sources.
- Adopt a common and shared coding according to the ***ontology of the analysis domain*** and that has been treated with ***data cleaning*** and ***data imputation algorithms*** that have eliminated inconsistencies and missing data in the variables of interest.

An automated data pre-processing strategy can be studied and implemented to homogenise the record layouts of the various simulators and of the various databases, defining a common coding, identifying a series of consistent rules which allow for assessing the quality of the data and developing a data imputation and data fusion procedure crucial for the cleansing of the final dataset. The innovative strategy of data editing suited for the i4Driving Project starts from the contributions of the TREEVAL data editing procedure (Petrakos, et al., 2004) and the new paradigm of missing data imputation and data fusion coined by D'Ambrosio et al. (2012). A need for Data Fusion methods could arise while copying with ***Naturalistic Driving Study Data*** (Guo, 2019).

How to choose the Data Mining method within a specific application scenario?

"Broadly speaking, data mining is all about finding summary relationships between the entries in the data matrix that are either unusually frequent or unusually infrequent." (Aggarwal, 2015). The data matrix can be analysed with the perspective of either the rows of the data matrix (namely the data points), or the columns (namely the variables). Regarding the columns, we are interested in analysing the relationships among the variables to study correlation, association, to summarise their variation and reduce their dimensionality. Regarding the rows, we are interested in partitioning the data points into homogeneous groups characterised by similar measurements of the variables.

Data Mining, in the core phase of modelling for data analysis, makes use of Statistical Learning. **Statistical Learning** refers to a vast set of methods for understanding *"what the data says"* (Hastie, Tibshirani, Friedman, 2011, 2009). These methods can be classified into unsupervised learning and supervised learning.

- **Unsupervised learning** moves in the discovery context to detect the interesting patterns, groups, anomalies, associations, correlation, similarities, typologies, clusters. This yields to ***Exploratory Data Analysis***.
- **Supervised learning** moves in the confirmatory context to identify the functional relationship between a response or target variable and a set of predictors for ***Modelling and Prediction***.

The main distinction between unsupervised and supervised learning is in the role played by the variables. All variables play the same role in unsupervised learning, whereas there is one target variable to be explained and/or predicted based on a set of other variables in supervised learning.

As an example, in a dataset of driving events including a set of variables related to the driver, the vehicle, the trip, the road, etc. the interest is in detecting the most interesting patterns of events, the typologies of drivers, clusters of driving events, each characterised by different typologies of drivers, trips, and vehicles. In the case where a label class is associated to each event of a target or response

variable, specifying the follow-up of the driving event into either a crash or not, the interest is in modelling and predicting the probability that certain driving events result in a crash and then understanding which are the most common predictors. In this case, there is a need for supervised learning.

The target variable supervises the learning process. When the target is numerical, we deal with a **regression problem**. When the target is categorical, we deal with a **classification problem**. The binary variable "crash" or "not crash" helps us to supervise the learning process to identify among the predictors which are the key characteristics to discriminate a "good" driving event (yielding to "not crash") from a "bad" one (yielding to "crash"). This is an inferential task. We can also have a prediction task. We can use the fitted model to predict the probability of "crash" such to be able to assign the target class to a new driving event based on the predictors' measurements.

Many Data Mining methodologies have been exploited to gain knowledge of driver behaviour in Naturalistic Driving Studies (Murphey *et al.*, 2020) as well as investigating the patterns of road accidents (Montella *et a*l., 2020).

Application Scenarios in the **i4Driving domain** thus require:

- **Exploratory Data Analysis** to discover patterns, associations, co-occurrences of facts, similarities, typologies, correlation, clusters, etc.
- **Modelling and Prediction** to identify the functional relationship between a target variable and a set of predictors, which helps in decision-making and scenario analysis.

Data mining methods can be catalogued with respect to the perspective view of the data matrix to perform the data analysis and the type of learning. This yields to **a set of fundamental methodological building blocks for the i4Driving Project:**

- **Unsupervised Learning for Exploratory Data Analysis**
  - ✓ Association Rules Discovery for sparse binary databases
  - ✓ Factorial Methods for dimensionality data reduction and visualization
  - ✓ Clustering Methods for pattern recognition
- **Supervised Learning for Modelling and Prediction**
  - ✓ Interpretable Models for Regression and Classification
  - ✓ Black-Box Models for Regression and Classification

Anomaly/Outlier Detection Analysis can be approached using both unsupervised and supervised learning.

Preference Learning for the analysis of rankings data require suitable non-standard data mining methods (D'Ambrosio et al., 2017a).

2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) in the discovery context is done to extract useful information from data. Typically, it considers an unsupervised learning approach where all variables play the same role in the analysis. The main characteristic of EDA is that it does not consider probability assumptions. The aim is to detect the interesting patterns, groups, anomalies, associations, correlation, similarities, typologies, clusters, etc in the data. In the following sections, the methodological building blocks for EDA are discussed, namely **Association Rules Mining, Dimensionality Data Reduction, Clustering Methods.**

2.1 Association Rules Mining

Association Rules Discovery can be very useful in the **i4Driving domain**, analysing the co-occurrence of factors such as darkness and speed in accidents with a dead person. Considering that the percentage of accidents with a dead person is very low, it can be considered as an ***anomaly detection case***.

Let us consider the dataset of accident events where all information about the event has been recorded, such as the type of accident (out of road, crash against another vehicle, etc.), the type of road (highway, extra-urban, urban, etc.), the weather condition (rain, fuggy, serene, etc.), the light (day or night), the roadbed (dry, wet), whether there is a dead person or not, etc. The challenging question is to detect ***a set of typologies of accident events***, each with its main characteristics. This can be useful to improve the road signs, to add safety sensors in the car, etc.

All features can be converted into binary variables - called items - so that the matrix can be rather sparse.

Ideally, many two-way and three-way cross-classifications could be made to discover co-occurrences of two or three items. However, one then needs to consider how to process all possible combinations of items. A sophisticated methodology is required to deal with a big and sparse data matrix. In one accident event, there are very few items with respect to all possible items that can occur in any accident event. An association rule is a statement like ***<< "Night light" implies accident type "Out of road">>.***

Two key concepts can be applied. One is the ***Support***, to say how many times one item occurs, as well as two or three or a subset of items occur together. Another one is the ***Confidence,*** to say how many times "Out of road" among those with "Night light". If this percentage was the same of all item sets with just "Out of road", then there is no association between "out of road" and "Night light". The gap as measured by the ***Lift*** will tell us about the strength of the association rule << "Night light" implies accident type "Out of road">>.

This takes place in two steps. The first step is to discover a set of items that occur frequently together in a dataset, the so-called frequent items: these are rules with a 'Support' bigger than a fixed minimum level. This is the most difficult step, requiring a selective algorithm. The A-Priori Algorithm is the most used. The second step is to identify as Association Rules those with a 'Confidence' higher than a fixed threshold. In a 'Crash' Scenario, ***Association Rules Mining*** has been considered ***in combination with Classification and Regression Trees*** to analyse the powered two-wheeler crashes in Italy (Montella, 2011, Montella et al., 2011b, 2012).

2.2 Dimensionality Data Reduction

Dimensionality Data Reduction is a fundamental step of Exploratory Data Analysis and means that we need to reduce the dimensionality of data, thus reducing the number of variables to be considered

without losing the information they provide. A subset feature selection can be performed based on the challenging questions to be satisfied, as well as the quality of data.

The real mission is to provide the feature summarisation, removing the redundancy of information due to highly correlated or associated variables. This can be achieved by a factorial method with a suitable data visualization and analysis of the results.

The fundamental factorial method is **Principal Component Analysis** when all variables are numerical. Different approaches include the following:

Correspondence Analysis is a factorial method for the analysis of two-way contingency tables, while **Multiple Correspondence Analysis** deals with multidimensional qualitative data.

Non-linear factorial methods can also be performed considering the optimisation criterion based on Alternating Least Squares which considers the nonlinear transformation of the variables. **Exploratory Projection Pursuit** is another approach to explore non-linearity in multidimensional data.

Dimensionality Data Reduction and Visualisation Methods of Data Mining include **Procrustes Analysis** and **Multidimensional Scaling Methods**.

These are a set of procedures for the analysis of one or more matrices of "proximity measures" among all the possible pairs of objects. The aim is to obtain a geometric configuration of the objects in a reduced number of dimensions, say two or three, such to reflect the hidden structure of the data, in the sense that the greater the similarity between two objects, the closer they should be in the factorial configuration.

Scaling techniques are often applied in the social and behavioural sciences to study the relationship between objects and people which, in psychometric terms, are called "stimuli". These can occur in driving behaviour analysis.

2.3 Clustering Methods

Cluster Analysis does an unsupervised partitioning of data points into groups called clusters, which include very similar data points. Similarity is measured in terms of all features measurements and not just of one target variable as in supervised classification.

The aim is to discover what similar data points have in common, thus which features of measurements characterise one cluster with respect to another. Any cluster becomes a typology or a segment. Fundamental is the choice of the proximity measure such as the similarity or dissimilarity measure for categorical data and the distance measure for numerical data to evaluate how similar or different are two data points considering all features measurements.

There are many application scenarios of cluster analysis for **i4Driving domain.** The clustering method is used for Exploratory Data Analysis to provide Data Summarisation and concise insights from the data. It can be used for Driver Segmentation or collaborative filtering, in which the stated or derived preferences of a similar group of drivers is used to make driving recommendations within the group.

Cluster Analysis might also be considered for **Anomaly Detection**. In ***Driving Simulator Studies***, Cluster Analysis has been carried out to identify homogeneous drivers' average speed profiles in relation to different design alternatives. Grouping similar driving behaviours has allowed to identify the key discriminant factors characterising these behaviours with respect to specific driving conditions (Galante et al., 2010, Montella et al. 2011). Similar study has allowed to analyse the lateral position profiles.

We can distinguish between **hierarchical clustering methods** and **non-hierarchical clustering methods**.

When dealing with time series, these can be very noisy and sparse so that an appropriate model describing them can be hard to define. A non-hierarchical and flexible model-based approach can be fruitfully considered, such as ***Parsimonious Time Series Clustering using P-Splines*** (Iorio et al., 2016).

Non-Hierarchical Clustering can also be distinguished into Hard and Soft Clustering Approach.

In **hard or crisp clustering,** each data point is deterministically assigned to a particular cluster.

In **soft clustering**, each data point may have assigned a probability or likelihood or membership degree to many (typically all) clusters. Soft Clustering adopts the fuzzy logic, sometimes these algorithms are known as fuzzy clustering. Probabilistic Distance-based Clustering such as ***Boosted-oriented probabilistic smoothing-spline clustering*** can be proposed for multidimensional time series (Iorio et al., 2022).

The clusters may be hard to model with a prototypical shape implied by a distance function or probability distribution. Main motivation to perform a ***Density-Based and Graph-Based Clustering Algorithms*** is to find clusters with arbitrary shape. The key concept is that clusters are dense regions of objects in the data space that are separated by regions of low density representing the noise. Density-Based Clustering includes Grid-Based Methods, Density-based spatial clustering of applications with noise, known as DBSCAN algorithm, DENsity CLUstering or DENCLUE Algorithm.

3. Modelling and Prediction

The Statistical Learning Theory (SLT) paradigm (Vapnik, 1995, 1998) and its decline in machine learning models for inference and prediction (Hastie et al., 2009; James et al. al., 2021) are discussed considering the condition of application, the model assessment, the model selection criteria below.

3.1 Supervised Learning Methods

In Supervised Learning, there is a target variable (also known as response variable or output) which supervises the learning process and a set of predictors (also known as input). As mentioned above, when the target is numerical, we deal with a ***regression problem***. When the target is categorical with a set of response classes, we deal with a ***classification problem***. Two are the main goals:

- **Inference**, to identify the best model to understand or explain how the response variable depends on the predictors.
- **Prediction**, to assign a response class/value to a new data for that only the predictors' measurements are known.

Supervised Learning requires the formulation of a **statistical model** which is built up of two components:

- The ***systematic component***, which is the expectation of the response given the predictors' measurements, and
- The ***error component***, which is a random variable with zero mean that describes the *intrinsic randomness* underlying any statistical measurement.

A distinction is made between regression models when the response is numerical and classification models when it is binary or multinomial. As an example, when the response is binary of Bernoulli type, the systematic component is the probability of success.

The choice of the statistical model means specifying the systematic component.

A training sample is used to fit the model and to verify if the sample is coherent with the model specification. An independent test sample can be used to estimate the prediction error. Resampling methods (Efron, 1978, 1982) can be used, specifically ***bootstrap for model assessment*** and ***cross-validation for model selection***.

In the Statistical Learning Theory of Vapnik (1995, 1998), the goal is to identify the **learning machine** characterised by the best functional relationships between the input and the output such to approximate the supervisor's response minimising the loss of discrepancy or error ***(transductive inference)***. During the training, the learning machine constructs some operator which can be used for prediction of the supervisor's answer of any specific input vector. Selecting the best approach is the big challenge of statistical learning in practice (Hastie, Tibshirani, Friedman, 2001, 2009).

3.2 The bias-variance dilemma

Complex models lead to a phenomenon known as ***overfitting***, which essentially means the model follows the errors, or noise, too closely. We are over training the learning process, and the learning machine cannot be generalised to new data. Models which are too complex model result in ***propagation error,*** as the model would be erroneously generalised to fresh data. This is known as the ***trade-off bias-variance*** dilemma (figure 5):

- ***Variance*** refers to the amount by which the fitted model would change if we estimated it using a different training sample (Propagation Error).
- ***Bias*** refers to the error due to approximating a real-life problem, which may be extremely complex, by a much simpler model (Model Inadequacy Error).

As the model complexity of the learning machine increases, the variance tends to increase, and the squared bias tends to decrease (and vice versa).

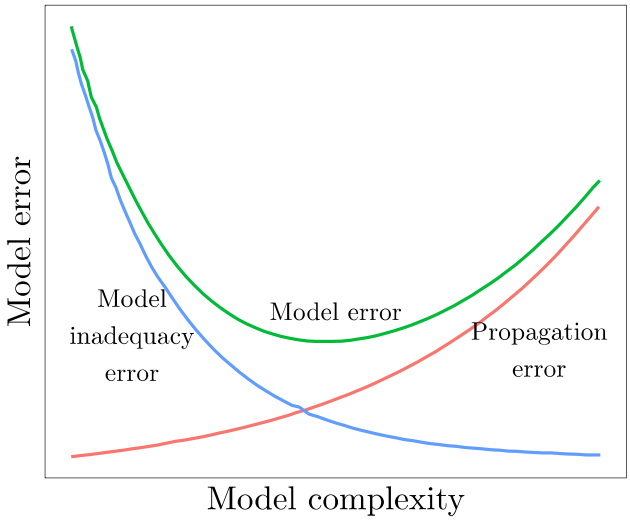

Figure 5. Trade-off Bias-Variance.

To minimise the expected prediction error, we need to select a learning method that simultaneously achieves low variance and low bias. If the expected prediction error tends to zero, the generalisation error tends to the variance of the error term of the statistical model, which is the intrinsic randomness of the induction approach, the so-called irreducible error. The first goal of any supervised learning method is to avoid the overfitting phenomenon.

3.3 Interpretable versus Black Box Machine Learning

Model selection and assessing the model accuracy for prediction of "fresh" data is the key point in machine learning.

There is another trade-off regarding ***interpretability versus flexibility*** in the choice of the statistical/machine learning model (figure 6).

Assuming a parametric form such as linear regression simplifies both interpretation and estimation, but it can be far from the true model that has generated our sample data and our estimate can be poor. A too simple model yields to ***model inadequacy error*** (figure 5).

Parametric modelling when the size of data is not huge includes ***linear regression, logistic regression and discriminant analysis.*** Dealing with high dimensionality data, it might be convenient to use ***penalised regression*** such as ***LASSO, RIDGE, ELASTIC-NET Regression*** or dimension reduction methods such as ***Principal Component Regression, Partial Least Squares Regression.***

Choosing flexible models based on **non-parametric methods** fit many different functional forms requiring the estimation of a higher number of parameters; thus they require huge size. The naïve learning machine is the ***K-Nearest Neighbour***.

Flexible models allowing to deal with non-linearity are represented by non-parametric models, such as ***Regression Splines, Smoothing Splines and Local Regression***.

There is also a **semi-parametric modelling** approach provided by ***Additive Models, Partially Linear Models, Generalized Linear Models, Generalized Partial Linear Models.***

Non-parametric methods which have a distribution-free approach and ***interpretable learning machines*** are ***Classification and Regression Trees, Decision Trees*** for both standard data (Breiman et al., 1984; Mola and Siciliano, 1997, Iorio et al. 2019) and non-standard data (Siciliano and Mola, 2000, D'Ambrosio et al., 2017b)**.** Their flexible versions to improve the prediction accuracy are given by ***Ensemble Methods*** (Dietterich, 2000), such as ***Bagging*** (Breiman, 1996), ***Boosting*** (Freud and Shapire, 1999) ***Random Forest*** (Breiman, 2001).

Support Vector Machines and ***Projection Pursuit Regression*** are also flexible methods.

Deep Learning with ***Neural Networks*** (Aggarwal, 2018) are black box machine learning models which guarantee good performance in terms of prediction accuracy. They represent a fundamental approach to deal with big data and non-standard data types**.** ***Convolutional Neural Networks*** are ideally for image classification, ***recurrent neural networks*** for time series and other sequences.

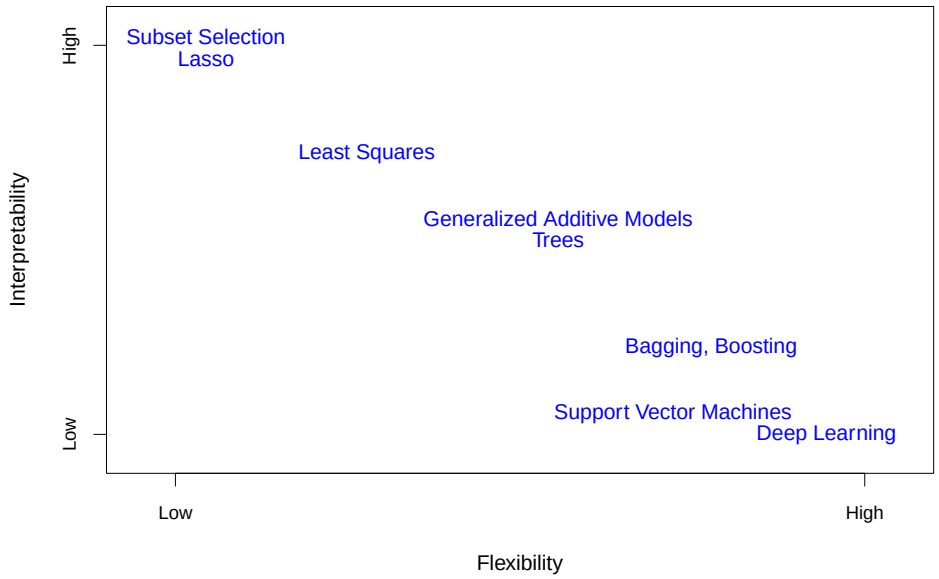

Figure 6. Trade-off Interpretability-Flexibility (James, Witten, Hastie & Tibshirani, 2009).

In **Instrumented Vehicle Studies**, ***non-parametric regression using smoothing fit*** has been applied to identify the driving behaviour and which factors influence its variations, to analyse the effects of an advanced driver-assistance system to improve safety of cyclists overtaking (Rella Riccardi et al., 2022a, b, c).

In **Crash Scenarios, *Classification Trees,*** and ***Association Rules Discovery*** were considered to analyse the powered two-wheeler crashes in Italy (Montella et al., 2011, 2012). ***Random Forests, Support Vector Machines, and Neural Networks*** were considered for ***Crash Scenario Prediction*** (Rella Riccardi et al., 2022).

In the realm of supervised methods, relevance should be attached to approaches for modelling Naturalistic Driving Study Data, for instance to obtain a model for ***risk prediction*** based on the

characteristics of the driver (Guo & Fang, 2013), including his/her behaviour as well as some data about the road context and the vehicle.

The use of ***ML classifiers***, like ***Naïve Bayes, K-Nearest Neighbour (KNN), AdaBoost, Random Forest***, and ***Support Vector Machine***, is widespread in the field of ***Road Traffic Accident Analysis*** (Bokaba *et al*., 2022).

Some classification models (***KNN, SVM***, and ***DNN***) can fruitfully be used to investigate ***human-vehicle interaction***, e.g., to recognise the ***driving risk status*** (Wu *et al*., 2023).

Even if a large class of modern ML models is less interpretable than classical ones, beside a good accuracy, they can still offer useful knowledge on the phenomenon. For instance, ***Random Forest*** can be used ***to rank the importance of features of traffic accidents*** that are more relevant in the light of a chosen outcome (Yang *et al*., 2023).

Furthermore, ***Deep Learning*** approaches are customary to ***model driver attention*** (Gao & Murphey, 2023).

4. Sensitivity Analysis

Sensitivity analysis (SA) assesses how much the uncertainty of a model output depends upon its inputs. Though it is generally agreed in existing guidelines that uncertainty and sensitivity analyses are both crucial for the validation or verification of a model, their application is hampered by practical difficulties, scarce awareness and at times a reluctance to expose the weaknesses of a model.

We present here global sensitivity analysis (GSA), mainly through one class of global SA methods known as 'variance-based' methods – considered by most practitioners as a recommended practice – and offer pointers on additional methods. We also suggest several hints for a successful and effective use of these techniques.

4.1 Uncertainty versus Sensitivity Analysis

It is important to define to key terms, namely:

- **Uncertainty analysis** (UA): The quantification of the uncertainty in model output; and
- **Sensitivity analysis**: The study of the relative importance of different input factors on the model output uncertainty.

As we shall discuss here, the two analyses are linked. To a natural scientist trained in calculus, sensitivity analysis may evoke the derivative of a function of interest with respect to its inputs.

In most ecological studies, the factors may vary considerably, from a few percent to orders of magnitude, and likewise the output because of error propagation. Hence, to an ecologist, what happens to y in a single point of the multidimensional space of existence of $x_1, x_2, \ldots x_k$ may be uninformative; ecologists will want sensitivity measures that are global, i.e., concerned with the whole space of variability of the inputs.

When the overall uncertainty in y is modest, it is not so important to ascertain where this is coming from. Conversely, if y spans orders of magnitude, then SA becomes indispensable to understand the system studied and pinpoint the factor(s) that convey the most uncertainty. Such information might help to guide further research by highlighting where efforts on data collection should focus to maximise the reduction of uncertainty in the output.

4.2 Variance-based sensitivity analysis

The starting point for UA is the analytic or computer-coded form of the model and the probability distributions of the inputs. Although determining these probability distributions is preliminary to any analysis, it is often the most important and expensive part of the work. This stage of elicitation may involve experts from several disciplines and/or the collection of a considerable amount of data.

Monte Carlo based UA consists of a series of simulations. In each of these simulations, the value of each input factor is sampled from its distribution. The corresponding output value produced is recorded and the statistical properties of the output distribution are finally analysed.

We now concisely describe the SA variance-based methods. A handbook is available for a more detailed treatment of the topic (Saltelli et al. 2008). These measures are mostly due to the work of Russian mathematician Ilya M Sobol (Sobol, 1993). We take the variance of the output as the target of the analysis, and following statistical theory, we decompose it following the ANOVA (ANalysis Of VAriance) scheme to obtain the sensitivity coefficients S, the objective of the estimation procedures of variance

based SA methods. Without proof, we add here a few notes: the first order term S_i is identical to the Pearson correlation ratio η_i^2: in other words, S_i computes the mean of a moving average, see Figure 7.

Monte Carlo estimation of both indices is straightforward and based on a single Monte Carlo loop. These formulae can be found in Saltelli *et al.* (2008), and a discussion is in Saltelli *et al.* (2010). For most applications in ecology, it is sufficient to compute the S_i and the T_i. If all S_i's are equal to the corresponding T_i's the model is said to be additive, i.e., without interactions. This information is considerably superior to that offered by the derivatives because it captures interactions among factors.

Variance based methods offer the advantages to be grounded in statistical theory; to decompose the variance into sets of factors; they are easy to interpret, i.e. the total sensitivity index is the fractional variance that would be left on average if all factors but x_i could be fixed; S_i's and T_i's can be linked to well-defined experimental settings – e.g. in order to decide if a variable can be fixed one need to use T_i and not S_i. All this is treated in Saltelli *et al.* (2008).

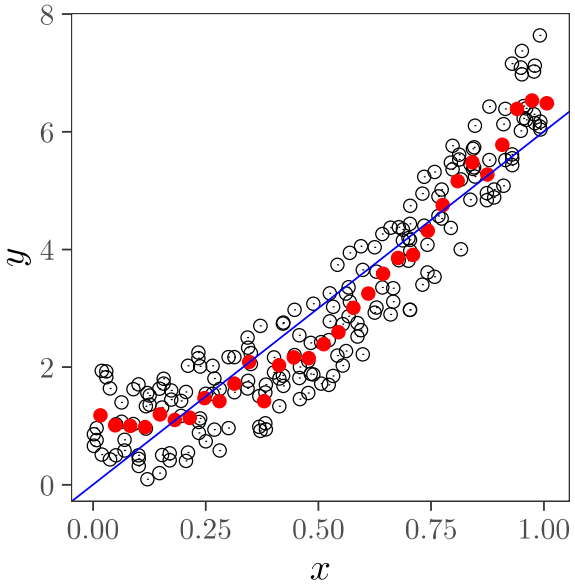

Figure 7. Scatterplots with moving averages (red). The straight line is the standardized regression coefficient of y on x_i, the discontinuous line is the moving average E_(x_(~i)) (y ⎪ x_i) (Saltelli, 2008).

4.3 Sensitivity analysis in practice

Here we provide a series of suggestions for the practitioner:

Choose one and only one output of interest. Since a model may produce many outputs (e.g., time-series, spatially distributed), we suggest running SA only on the output that helps to answer the question posed by the analysis.

Be open to the possibility that the model produces uncertainties so wide as to make its predictions irrelevant. If this happens, it could simply mean that the quality of the evidence feeding into the model does not allow meaningful estimates to be produced. One should then change the model, or the

question asked from it. We recommend this approach to tame 'modelling hubris', e.g., the temptation to develop larger and larger models (Saltelli et al. 2020), see Figure 5.

Consider extending the set of input factors using triggers. If one is uncertain about epistemic features of the model – e.g., what formula to use for a particular phenomenon in the model, a trigger may allow one to select two or more formulae 'at runtime' – e.g., if $x_i < 0.5$, then choose formula A, if $x_i \geq 0.5$ use formula B. The same may apply to different grid resolutions, choice of algorithms in the model, and so on. The effect of triggers on the model output should be examined jointly with parametric uncertainties to capture possible interaction effects.

Why run a model just once? In the process of building a model, time and effort can be minimised by running systematically the model in Monte Carlo simulations: instead of executing the model once, execute it one hundred times, or even maybe only ten. Interesting discoveries or questions may arise:

- Bugs can be detected more quickly and fixed, instead of carrying them forward in the model building.
- An addition to a model makes no change to the output in none of the points tested; is the addition necessary?
- An addition makes a change which exceeds expectations; why was this the case?

Avoid lying with SA. One can lie with SA by varying only some factors, implicitly assuming that all others are perfectly known. In an adversarial setting, this risk being exposed by the opposing party. Scarce attention to uncertainties ultimately erodes trust in modelling. It happens frequently that models run to produce point estimate are revealed as non-conservative when uncertainties are properly plugged in (Puy, Lo Piano, and Saltelli 2020). An OAT approach is also vulnerable to deconstruction for the reasons discussed above. Another way of making a perfunctory SA is to bypass the stage of careful appraisal of the $p_1(x_1), p_2(x_2), \ldots p_k(x_k)$ and perform an analysis where all factors have the same uncertainty, e.g., 5% or 10%. These analyses are a case of GIGO, garbage in, garbage out, as instances where all factors are equally uncertain are possibly non-existent in ecology.

Consider via negativa. Some authors, including us, recommend using models also to disprove rather than to prove a give thesis (Oreskes 2010; Oreskes, Shrader-Frechette, and Belitz 1994; Saltelli and Giampietro 2017). Via negativa can provide valuable insights because:

- 'Wrongs' are more evident than 'rights'.
- Knowledge grows by subtracting what cannot be.
- "Actions that remove are more robust than those that add because addition may have unseen, complicated feedback loops" (Taleb 2012).

Other methods. The literature offers several other interesting methods for SA. When for some reason one is not interested in the variance of the output, e.g., because its distribution is very skewed or long-tailed, then one may resort to moment-independent measures. These permit e.g., ranking factors based on how – fixing them – affects the entire probability distribution function – rather than just its variance. These measures are named moment-independent (Borgonovo and Iooss 2016).

Shapley coefficients used by economists can be related to the sensitivity coefficients just discussed (Owen 2014). Many practitioners use the method of Morris (Morris 1991), which is also close to the total sensitivity index T_i and is recommended when only few simulations can be performed. Morris needs more modelling assumptions than T_i and is more cumbersome to interpret as it produces two measures for each factor. For this, we would rather suggest T_i at low sample size rather than Morris (Campolongo, Saltelli, and Cariboni 2011).

Large, CPU-intensive models: Variance-based indices are rather expensive to compute in terms of number of simulations; computing all the S_i's and all the T_i's may come to a cost of $N(k+2)$ where N may be of the order of hundreds or thousands. When the model cannot afford this number, one may use emulators, replacement models that run cheaply. See for an example (Schöbi, Sudret, and Wiart 2015).

Other readings: Razavi *et al.* (2021) describes future orientations for SA. Recent reviews are Norton (2015) and Wei *et al.* (2015).

4.4 Integrate GSA with machine learning algorithms

Based on existing experience in the consortium on data mining in the context of accident analysis (Montella et al. 2012) and on GSA (Saltelli et al. 2008; 2021; Puy et al. 2022) a combination of machine learning techniques is being developed for the analysis. These include clustering techniques and (generalised) linear mixed effect models, feature selection via regularised regression tools (e.g., LASSO, elastic net), Random Forests and subset selections. Additionally, GSA methods can be used for selecting features, following recent developments on the use of SA of and for data mining (Tunkiel, Sui, and Wiktorski 2020; Antoniadis, Lambert-Lacroix, and Poggi 2021), including using the concept of mean dimension (Hoyt and Owen 2021). Note also section 3.4 of a recent position paper on sensitivity analysis (Razavi et al. 2021). Interesting as a possible linkage sensitivity analysis – sensitivity auditing also this paper (Bénesse et al. 2022, Piano et al. 2022).

Specifically, i4Driving will test via machine learning the total sensitivity indices (Homma and Saltelli 1996) that have already found use in an adjacent field – model selection in regression (Becker, Paruolo, and Saltelli 2021). In addition, the use of sensitivity analysis as a contribution to model interpretability is also considered (Iooss, Kenett, and Secchi 2022). Additional avenues for research are the use of pre-integration techniques for SA (Liu and Owen 2022) based on Paul Constantine's active subspace decomposition (Constantine, Dow, and Wang 2014).

Moreover, a new measure of SA based on the concept of discrepancy (Puy, Roy, and Saltelli 2023) has been developed to be used for machine learning investigation of i4Driving. While SA improves the transparency and reliability of mathematical models, its uptake by modelers is still scarce. This is partially explained by its technical requirements, which may be hard to understand and implement by the non-specialist. An approach of SA based on the concept of discrepancy that is as easy to understand as the visual inspection of input-output scatterplots can be adopted in practical problems. Some discrepancy measures can rank the most influential parameters of a model almost as accurately as the variance-based total sensitivity index. Moreover, it can be used an ersatz-discrepancy whose performance as a sensitivity measure matches that of the best-performing discrepancy algorithms, and it is simple to implement, easier to interpret and orders of magnitude faster.

Finally, the project i4Driving can help in understanding the union of qualitative and quantitative aspects of uncertainty: it is acknowledged in some recent work on sociology of quantification (Di Fiore et al. 2022) and on impact assessment (Saltelli et al. 2023).

4.5 Sensitivity auditing

Sensitivity auditing is an extension of SA to include in the analysis the entire model building process, with emphasis on bias, motivations and expectations of both users and developers (Saltelli et al., 2013). It will be used in i4Driving modelling studies. Applications of sensitivity auditing are described in Lo Piano et al. (2022, 2023).

5. Case Study on SHRP2 Databases

Statistical Learning and Data Analysis using Data Mining methods have been performed in a case study. The databases on mobility deriving from the **Strategic Research Program (SHRP2)** on Naturalistic Driving Study collected by University of Virginia have been considered (https://insight.shrp2nds.us). The aim is to address the driver performance and behaviour in traffic safety.

5.1 Exploratory Data Analysis of the "Event" dataset

5.1.1 Data Preparation

The Event dataset contains 41 530 rows (41 530 events) and 141 variables. For a detailed description of the variables, see the metadata description available on the website.

The target variable is the Event Severity (variable name: EventSeverity1, id variable: 14), of which the table 1 shows the frequency distribution.

	Level	N	%
EventSeverity1 (14)	Additional Baseline	12 581	30,294
	Balanced-Sample Baseline	19 998	48,153
	Crash	1 848	4,450
	Crash-Relevant	42	0,101
	Near-Crash	6 921	16,665
	Non-Subject Conflict	140	0,337
	TOTAL	**41 530**	**100,000**

Table 1. Frequency Distribution of the Event Severity.

The categories "Additional Baseline" and "Balanced-Sample Baseline" together sum up to 32 579 cases (about 78.45% of the total sample). Of the 141 variables, 36 contain exactly 32 579 missing values, as well as 66 variables contain more than 41 000 missing values. It is clear that if the outcome of an event is either "Additional Baseline" or "Balanced-Sample Baseline", many information is not collected, hence technically, we cannot interpret missing values as statistically missing values. The "complete" variables are 36, of which 9 are numerical (Event Start, FrntSeatPassngrs (Front Seat Passengers Details), RearSeatPassngrs (Rear Seat Passengers Details), SecTask1StartTime (Secondary Task 1 Start Time Details), SecTask1EndTime (Secondary Task 1 End Time Details), SecTask2StartTime (Secondary Task 2 Start Time Details), SecTask2EndTime (Secondary Task 2 End Time Details), SecTask3StartTime (Secondary Task 3 Start Time Details), SecTask3EndTime (Secondary Task 3 End Time Details)).

Except for the first variable, Even Start, for which the statistical descriptive is in table 2, the others either contain some coding for missing data (for example, the minimum of both FrntSeatPassngrs and RearSeatPassngrs is equal to -99) or they are not interesting for our purposes.

Mean	SD	Min	Q1	Median	Q3	Max
1017754,16	1432361,88	1	264179	585562	1184705	61316961

Table 2. Statistical Descriptive Summary of the Event Start.

The 27 categorical variables, together with the number of levels (categories) of each of them, are printed in table 3. We excluded from the analysis the variables "Driving Behaviour 2", "Driving Behaviour 3", "Sec Task 2" and "Sec Task 3" because they share the same categories, with the missing

one collapsed in either the category "unknown" or "other". Then, we merged some categories of the other variables. Precisely:

"Pre-Incident Manoeuvre": we combined the categories "Going straight, accelerating", "Going straight, but with unintentional drifting within lane or across lanes" and "Going straight, constant speed" into a single category *"Going straight (combined)".* Then we merged the categories with relative frequency lower than 0.5% ("Backing up (other than for parking purposes)", "Disabled or parked in travel lane", "Leaving a parking position", "Making U-turn", "Manoeuvring to avoid a pedestrian/pedal cyclist", "Manoeuvring to avoid a vehicle", "Manoeuvring to avoid an object", "Merging", "Other", "Stopped in traffic lane", "Unknown") into a single category named *"Other (combined)".*

Variable	**Pre-Incident Maneuver**	**Maneuver Judgment**	**Event Severity1**	**Driving Behavior1**	**Driving Behavior2**	**Driving Behavior3**
# Categories	22	5	6	56	58	45
Variable	**Impairments**	**Hands On Wheel**	**Driver Seatbelt**	**Light**	**Weather**	**Surface Condition**
# Categories	19	10	5	5	9	9
Variable	**Contiguous Travel Lanes**	**Through Travel Lanes**	**V1 Lane Occupied**	**Traffic Density**	**Traffic Control**	**Relation To Junction**
# Categories	10	9	15	8	15	10
Variable	**RdAlignment (Roadway alignment?)**	**Grade**	**Locality**	**Intersection Influence**	**Sec Task1**	**Sec Task2**
# Categories	5	5	11	8	64	58
Variable	**Sec Task3**	**Construction Zone**		**Traffic Flow**		
# Categories	44	4		5		

Table 3. List of Categorical Variables with their number of categories.

"Driving Behaviour 1": we merged the categories with relative frequency lower than 0.25%, namely those in table 4, into the category *"other (combined)".* Moreover, the categories "Improper turn, cut corner on left" and "Improper turn, cut corner on right" have been merged in the category *"improper turn (combined)".*

"Aggressive driving, other"	"Illegal passing"	"Parking in improper or dangerous location"
"Aggressive driving, specific, directed menacing actions"	"Improper backing, other"	"Passing on right"
"Apparent general inexperience driving"	"Improper signal"	"Right-of-way error in relation to other vehicle or person, apparent decision failure"
"Apparent unfamiliarity with roadway"	"Improper start from parked position"	"Right-of-way error in relation to other vehicle or person, other or unknown cause"
"Apparent unfamiliarity with vehicle"	"Improper turn, other"	"Signal violation, apparently did not see signal"
"Avoiding animal"	"Improper turn, wide left turn"	"Signal violation, intentionally disregarded signal"
"Avoiding other vehicle"	"Improper turn, wide right turn"	"Signal violation, tried to beat signal change"
"Avoiding pedestrian"	"Making turn from wrong lane"	"Speeding or other unsafe actions in work zone"
"Cutting in, too close behind other vehicle"	"Non-signed crossing violation"	"Stop sign violation, apparently did not see stop sign"
"Cutting in, too close in front of other vehicle"	"Other"	"Stop sign violation, intentionally ran stop sign at speed"

"Did not see other vehicle during lane change or merge"	"Other improper or unsafe passing"	"Sudden or improper braking"
"Disregarded officer or watchman"	"Other sign (e.g., Yield) violation, apparently did not see sign"	"Sudden or improper stopping on roadway"
"Driving in other vehicle's blind zone"	"Other sign (e.g., Yield) violation, intentionally disregarded"	"Unknown"
"Driving without lights or with insufficient lights"	"Other sign violation"	"Wrong side of road, not overtaking"

Table 4. List of categories to be merged into "other".

"Hands on Wheel": we merged the categories "Left hand at least", "Left hand off at least" and "Left hand only" into the category *"Left hand (combined)".* Then we merged the categories "Right hand at least", "Right hand off at least" and "Right hand only" into the category *"Right hand (combined)".* Finally, the categories "None", "None – Knees" and "Unknown" were combined into the category *"None (combined)".*

"Driver Seatbelt": the categories "None", "Unknown" and "Not applicable" were merged into *"None/Unknown(combined)"*

"Weather": the categories "Sleeting", "Snow/Sleet and Fog" and "Unknown" are merged in *"other(combined)".* Then, the categories "Fog" and "Rain and Fog" are merged in the category *"fog/rain_fog(combined)".*

"Surface Condition": the categories "Gravel over Asphalt", "Gravel/Dirt Road", "Icy", "Muddy", "Other" and "Unknown" are merged into *"other(combined)".*

"Contiguous Travel Lanes": the categories "7", "8+" and "unknown" are combined into *"7+".*

"Through Travel Lanes": the categories "5", "6", "7" and "8+" are combined into *"5+".*

"V1 Lane Occupied": the categories "5", "6", "7" are combined into *"5+".*

"Traffic Control": the categories "No passing signs", "Officer or watchman", "One-way Road or street" and "Other" are combined in "other(combined)". The categories "Railroad crossing with gate and signals", "Railroad crossing with markings or signs" and "Railroad crossing with signals" are merged in the category *"Railroad(combined)".*

"Relation To Junction": the categories "Other" and "Rail grade crossing" are combined in *"other(combined)".*

"RdAlignment": the categories "Other" and "Unknown" are combined in "other(combined)".

"Sec Task 1": the categories "Cell phone, Browsing", "Cell phone, Dialling hand-held", "Cell phone, Dialling hand-held using quick keys", "Cell phone, Dialling hands-free using voice-activated software", "Cell phone, holding", "Cell phone, Holding", "Cell phone, Locating/reaching/answering", "Cell phone, other", "Cell phone, Talking/listening, hand-held", "Cell phone, Texting", "Tablet device, Locating/reaching", "Tablet device, Operating", "Tablet device, Other" and "Tablet device, Viewing" are combined in *"Cell phone/tablet(combined)".* The categories "Child in adjacent seat - interaction" and "Child in rear seat - interaction" are merged in the category *"Child interaction(combined)".* The categories "Drinking from open container", "Drinking with lid and straw", "Drinking with lid, no straw" and "Drinking with straw, no lid" are merged in the category *"Drinking(combined)".* The categories "Adjusting/monitoring climate control", "Adjusting/monitoring other devices integral to vehicle", "Adjusting/monitoring radio" and "Inserting/retrieving CD (or similar)" are combined in the category *"Adjusting car devices(combined)".* The categories "Applying make-up", "Biting nails/cuticles",

"Brushing/flossing teeth", "Combing/brushing/fixing hair", "Reaching for personal body-related item" and "Shaving" are combined in the category *"Personal care(combined)".* The categories "Eating with utensils", "Eating without utensils" and "Reaching for food-related or drink-related item" are combined in the category *"Eating(combined)".* The categories "Extinguishing cigar/cigarette", "Lighting cigar/cigarette", "Reaching for cigar/cigarette" and "Smoking cigar/cigarette" are combined in the category *"Smoking(combined)".* The categories "Passenger in adjacent seat - interaction" and "Passenger in rear seat - interaction" are combined in the category *"Passenger Interaction(combined)".* The categories "Removing / adjusting clothing", "Removing/adjusting jewellery" and "Removing/inserting/ adjusting contact lenses or glasses" are combined in the category *"Removing/adjusting personal stuffs(combined)".* The categories "Distracted by construction", "Looking at an object external to the vehicle", "Looking at animal", "Looking at pedestrian", "Looking at previous crash or incident" and "Other external distraction" are merged in the category *"External distractions(combined)".* The categories "Insect in vehicle", "Object dropped by driver", "Pet in vehicle", "Reading", "Unknown", "Unknown type (secondary task present)", "Writing", and "Moving object in vehicle" are merged in the category *"Other(combined)".*

"Impairments": the categories "Angry;Drugs, alcohol;", "Drowsy, sleepy, asleep, fatigued;Drugs, alcohol;", "Drowsy, sleepy, asleep, fatigued;Other illicit drugs;", Other emotional state;Drugs, alcohol;", "Drugs, alcohol;", "Drugs, alcohol;Other illicit drugs;", "Other emotional state;Other illicit drugs;" and "Other illicit drugs;" are combined in *"Drug/alcohol(combined)".* The categories "Angry;" and "Angry;Other emotional state;" are combined in *"Angry(combined)".* The categories "Drowsy, sleepy, asleep, fatigued;", "Drowsy, sleepy, asleep, fatigued;Angry;" and "Drowsy, sleepy, asleep, fatigued; Other emotional state;" are combined in the category *"Drowsy(combined)".* The categories "Ill, blackout;", "Impaired due to previous injury;", "Other emotional state;", "Other;" and "Unknown;" are combined in *"other(combined)".*

5.1.2 Multiple Correspondence Analysis

Multiple Correspondence analysis (MCA) is a dimensionality data reduction method dealing with categorical data decomposed in factorial axes the Chi-Square inertia measure. The aim is to provide factorial representations visualising the association structure among the categories of the different variables and the patters of categories of the same variable. In this way, it is possible to identify typologies of driving behaviours. The MCA was performed by using all the selected variables except for "Event Severity 1", that has been used as a supplementary variable[3]. The dimensions are 142, their eigenvalues and percentage of explained inertia are summarised in table 5.

dimension	eigenvalue	percentage of inertia	cumulative percentage of inertia
1	0,25376085	3,752801362	3,752801362
2	0,19927977	2,947095243	6,699896605
3	0,15047647	2,225356309	8,925252914
4	0,13100998	1,937471476	10,86272439
5	0,10835858	1,602486084	12,46521047
6	0,09497191	1,404514137	13,86972461

[3] The supplementary variable, often the target variable, with its categories does not contribute to the geometrical identification of the factorial axes but it can be projected ex post using the analytical formulation of the factorial axes to analyse ex post which active categories are more related to the target response class by looking at their distance in the factorial representation.

dimension	eigenvalue	percentage of inertia	cumulative percentage of inertia
7	0,09379419	1,387097154	15,25682176
8	0,09160251	1,354685007	16,61150677
9	0,09023589	1,334474396	17,94598117
10	0,08335034	1,232645894	19,17862706
11	0,08155622	1,206113124	20,38474019
12	0,07899236	1,168196856	21,55293704
13	0,07829919	1,157945755	22,7108828
14	0,07784127	1,151173753	23,86205655
15	0,07342199	1,085818128	24,94787468
16	0,07150499	1,057468162	26,00534284
17	0,07000408	1,035271666	27,04061451
18	0,06920425	1,023443178	28,06405768
19	0,06694049	0,989965023	29,05402271
20	0,06393234	0,945478245	29,99950095
21	0,06317049	0,934211443	30,93371239
22	0,06051608	0,894956067	31,82866846
23	0,05964408	0,882060316	32,71072878
24	0,05931553	0,877201509	33,58793029
25	0,0586875	0,867913797	34,45584408
…	…	…	…
…	…	…	…
142	3,4296E-27	5,07201E-26	100

Table 5. Cumulative Inertia of Factors of Multiple Correspondence Analysis.

To better understand the phenomenon, as usual in dealing with MCA, the Benzecri correction has been applied. The result is shown in table 6.

dimension	Corrected inertia	Cumulative percentage of inertia
1	0,39432179	0,39432179
2	0,21503099	0,60935278
3	0,10023189	0,70958467
4	0,06651829	0,77610296
5	0,03595862	0,81206158
6	0,02228235	0,83434392
7	0,02123503	0,85557895
8	0,01935312	0,87493207
9	0,01822385	0,89315592
10	0,01305053	0,90620645
11	0,01184406	0,91805051
12	0,01022152	0,92827203
13	0,00980336	0,93807539
14	0,00953191	0,9476073
15	0,00710807	0,95471537
16	0,00616703	0,9608824

17	0,00547687	0,96635927
…	…	…
…	…	…
75	1,4097E-10	1

Table 6. Benzecrì Correction of Cumulative Inertia of Factors of Multiple Correspondence Analysis.

The first two dimensions explain more than 60% of the total inertia. Figure 8 shows the first factorial map (dimensions 1 and 2).

The first dimension (explaining more than 39% of the total inertia) discriminates between "Crash" and the other categories of the variable "Event Severity 1". Table 7 shows the most important categories in terms of their contribution to building the first and the second dimension.

Dimension 1		Dimension 2	
Categories	**Contribution**	**Categories**	**Contribution**
TrafficFlow_No lanes	14,697	ThrTrvLanes_1	8,255
ConTrvLanes_0	14,697	TrafficFlow_Not Divided(simple-2-way)	7,628
V1LaneOcc_No lane	14,467	Locality_Interstate	6,952
ThrTrvLanes_0	13,159	TrafficFlow_Divided	5,539
RelToJunc_Parking_within boundary	11,596	RelToJunc_Interchange area	5,015
PreInciManeuver_Entering a parking position	3,957	ConTrvLanes_2	4,545
IntersectionInfluence_Yes(Parking lot)	2,922	V1LaneOcc_1	3,896
Locality_Interstate	1,983	Locality_Moderate Residential	3,550
TrafficFlow_Divided	1,875	ThrTrvLanes_3	3,514
TrafficDensity_Level A1	1,734	TrafficFlow_No lanes	2,884

Table 7. Contribution of the most important categories to the factorial representation (first and second dimensions).

Hence, there is a good association between "No lanes" (Traffic flow), "0" (Contiguous Travel Lanes), "No lane" (V1 Lane Occupied), "0" (Through Travel Lanes), "Parking within boundary" (Relkation to Junction), "Entering a parking position" (Pre-incident manoeuvre), all placed on the right side of the plot, attracting the category "crash" to the same side. On the contrary, "Interstate" (Locality), "Divided" (Traffic flow) and "Level A1" (Traffic density) are all placed on the left of the figure. Note that, in the first dimension, all the other categories of the Event Severity are extremely close to the origin of the axis. The second dimension (about 21% of the inertia) is mainly characterized by the variables Locality ("Interstate", "Moderate Residential"), Traffic flow ("Divided", "Not divided (simple 2-way)", "No lanes"), and Through Travel Lanes ("1"). This dimension discriminates between the situations of "Near crash", "Crash relevant" and "Non subject conflict" and the others.

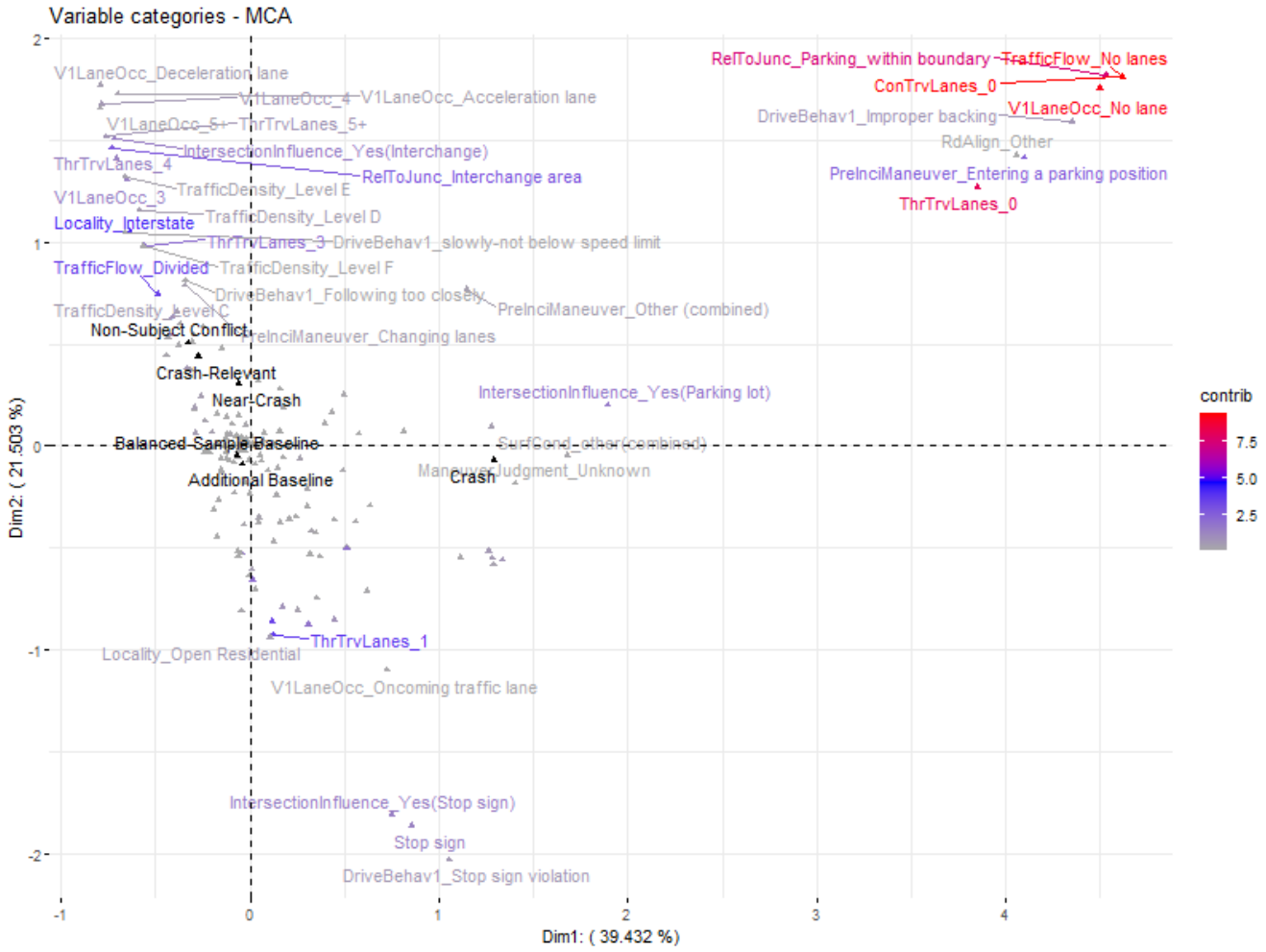

Figure 8. Factorial Representation of Multiple Correspondence Analysis (first and second dimensions).

Figure 9 shows the factorial map relative to the third and fourth dimensions. By looking at the plot, it can be noted that there is a clear difference between "crash", "near crash", "crash relevant" and "non subject conflict" and the other two categories. The variables that mainly contributes to both the dimensions are similar to the ones characterising the first two dimensions, with a difference in terms of the categories (table 8).

Dimension 3		Dimension 4	
Categories	**Contribution**	**Categories**	**Contribution**
IntersectionInfluence_Yes(Traffic sign)	11,679	IntersectionInfluence_Yes(Interchange)	10,045
Traffic signal	10,622	RelToJunc_Entrance/Exit ramp	9,797
V1LaneOcc_left turn lane	5,068	TrafficFlow_One-way-traffic	7,947
PreInciManeuver_Decelerating in traffic lane	4,637	ConTrvLanes_1	7,608
RelToJunc_Intersection-related	3,969	IntersectionInfluence_Yes(Stop sign)	3,568
Locality_Business	3,767	Stop sign	3,496
RelToJunc_Intersection	3,737	RdAlign_Curve right	3,297
IntersectionInfluence_No	3,593	TrafficFlow_Not Divided(center-2-way)	3,210
ConTrvLanes_2	3,051	PreInciManeuver_Negotiating a curve	3,105
V1LaneOcc_right turn lane	2,895	RelToJunc_Interchange area	2,535

Table 8. Contribution of the most important categories to the factorial representation (third and fourth dimensions).

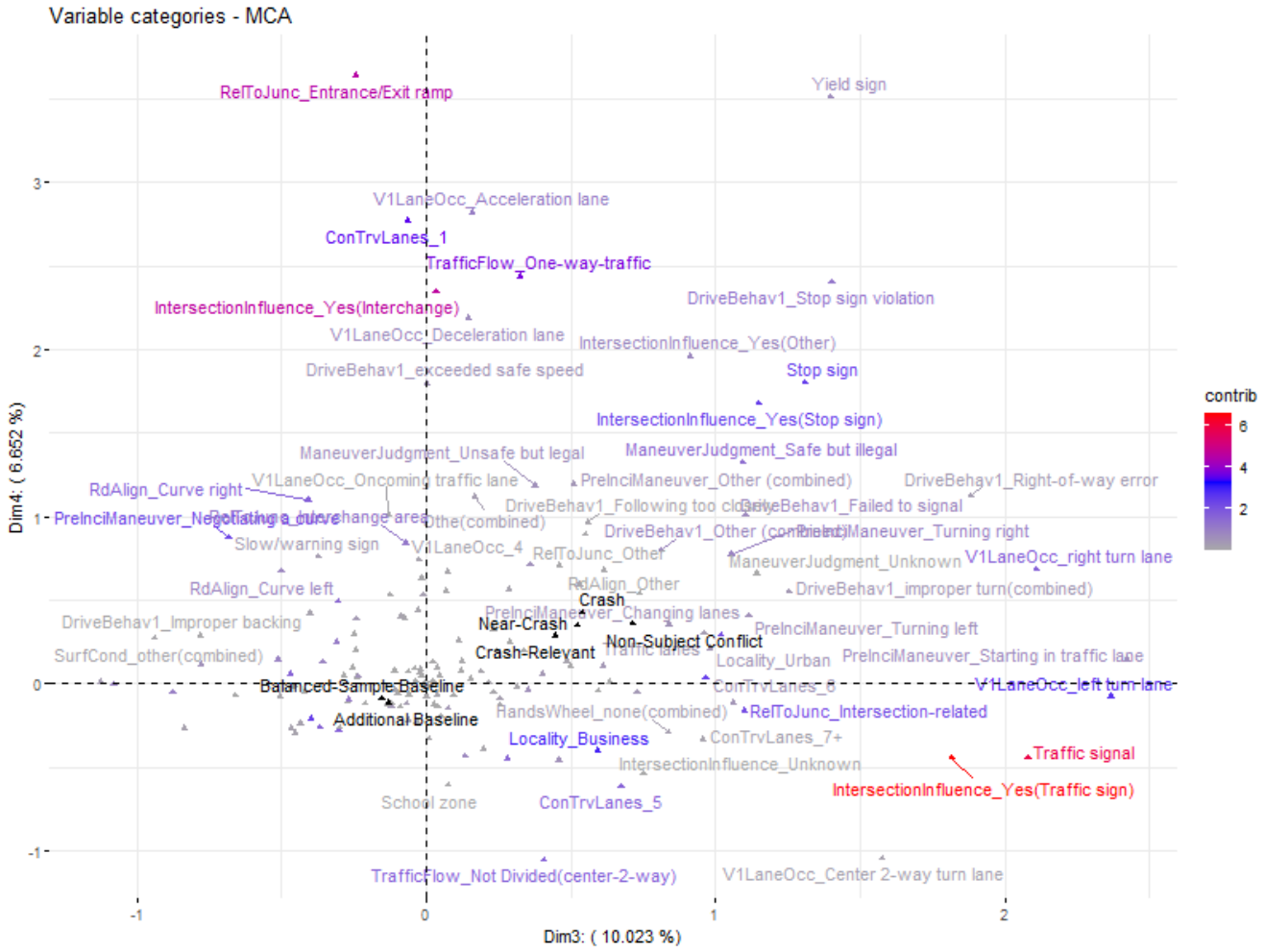

Figure 9. Factorial Representation of Multiple Correspondence Analysis (third and fourth dimensions).

The driving behaviour, some impairments, and the manoeuvre judgment characterise the third factorial map (dimensions 5 and 6) in figure 10. In this case, the sixth dimension (explaining about 2.2% of the total inertia) discriminate between crash-like events and the others due to drowsy driving behaviour, drowsy impairment, driving behaviour exceeding safe speed, unsafe manoeuvre judgment.

The table 9 shows the categories that mostly contributed to building dimensions 5 and 6.

Dimension 5		Dimension 6	
Categories	**Contribution**	**Categories**	**Contribution**
IntersectionInfluence_Yes(Stop sign)	11,349	Impairments_Drowsy(combined)	14,968
Stop sign	11,270	DriveBehav1_drowsy	14,324
TrafficFlow_One-way-traffic	7,672	ManeuverJudgment_Unsafe and illegal	6,887
DriveBehav1_Stop sign violation	7,469	ManeuverJudgment_Unsafe but legal	4,464
RelToJunc_Entrance/Exit ramp	7,085	IntersectionInfluence_Yes(Stop sign)	4,240
ConTrvLanes_1	5,206	DriveBehav1_Other (combined)	4,215
ManeuverJudgment_Safe but illegal	4,559	Stop sign	4,131
PreInciManeuver_Negotiating a curve	3,829	DriveBehav1_Exceeded speed limit	3,915
ManeuverJudgment_Unsafe and illegal	2,665	DriveBehav1_None	3,467
RdAlign_Curve right	2,351	DriveBehav1_exceeded safe speed	2,808

Table 9. Contribution of the most important categories to the factorial representation (fifth and sixth dimensions).

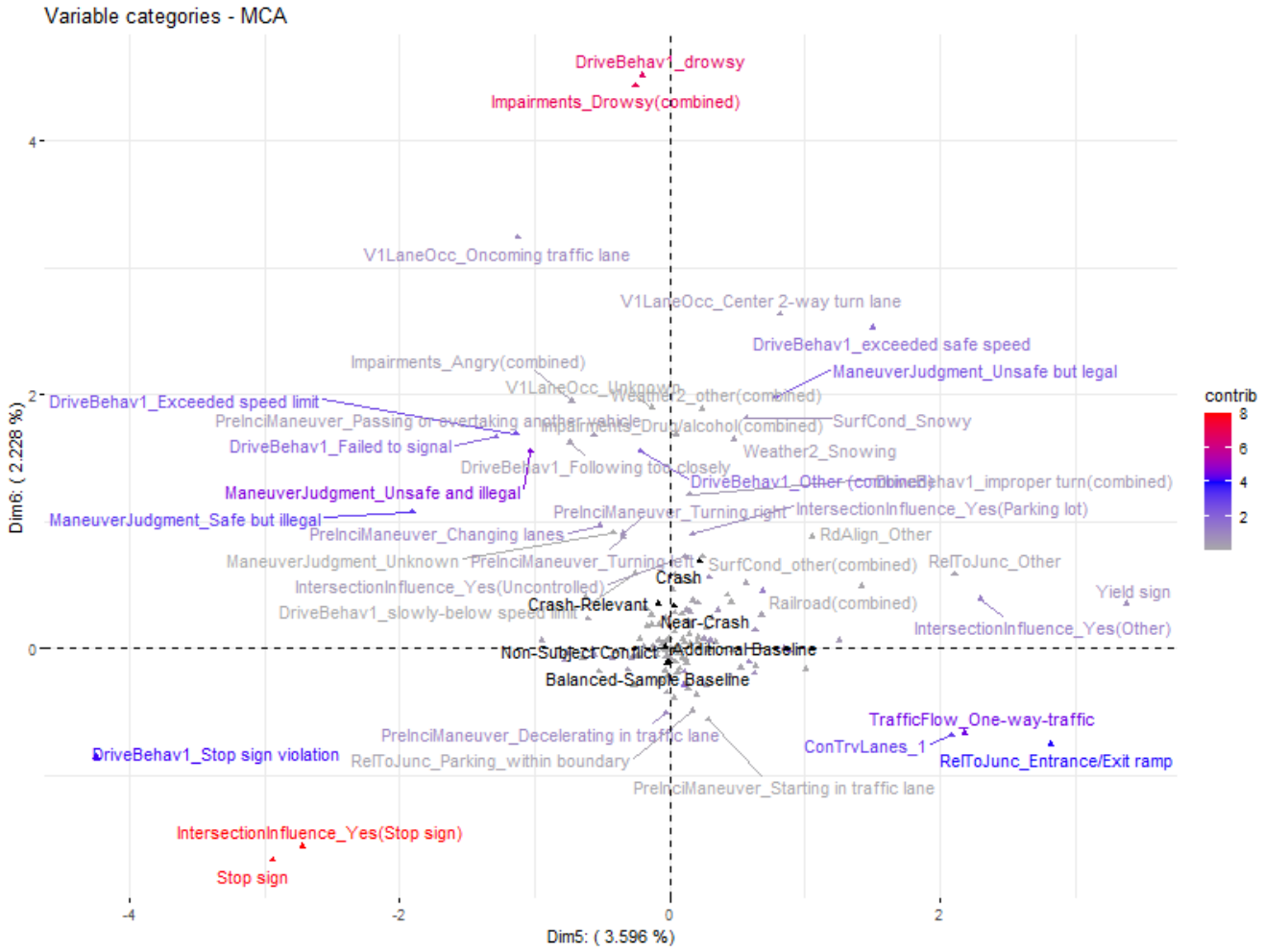

Figure 10. Factorial Representation of Multiple Correspondence Analysis (fifth and sixth dimensions).

5.2 Modelling and Prediction

5.2.1 Classification Trees

The classification tree has been built taking in account the subset excluding the events "Balanced-Sample Baseline" and "Additional Baseline" to focus our attention on relevant risk factors.

The subset counts 8 951 events. We used the same variables selected to perform the MCA. The prediction error of the tree is about 14.6%. *Twoing* was used as the splitting criterion. The figure 11 shows the classification tree. Each terminal node contains the frequency distribution of the response variable. In red the node numbers are reported.

The paths leading to each of the 9 terminal nodes with a given response class are:

- Node 7: **PreInciManeuver** ∈ {Entering a parking position, Other (combined), Turning left, Turning right} ∩ **DriveBehav1** ∈ {Improper backing, improper turn(combined)}, class: **crash.**
- Node 8: **PreInciManeuver** ∈ {Changing lanes, Decelerating in traffic lane, Going straight (combined), Negotiating a curve, Passing or overtaking another vehicle, Starting in traffic lane} ∩

DriveBehav1 ∉ {Improper backing, improper turn(combined)} ∩ **Traffic density** ∈ {Level A2, Level B, Level C, Level D, Level E}, class: **Near crash.**

- Node 9: **PreInciManeuver** ∈ {Changing lanes, decelerating in traffic lane, Going straight (combined), Negotiating a curve, Passing or overtaking another vehicle, Starting in traffic lane} ∩ **DriveBehav1** ∈ {Improper backing, improper turn(combined)} ∩ **Traffic density** ∈ {Level A2, Level B, Level C, Level D, Level E}, class: **crash.**
- Node 11: **PreInciManeuver** ∈ {Changing lanes, Decelerating in traffic lane, Going straight (combined), Negotiating a curve, Passing or overtaking another vehicle, Starting in traffic lane} ∩ **DriveBehav1** ∈ {Failed to signal, Following too closely, None, Right-of-way error, Stop sign violation, slowly-below speed limit} ∩ **Traffic density** ∈ {Level A1, Level F, Unknown}, class: **Near crash.**
- Node 13: **PreInciManeuver** ∈ {Entering a parking position, Other (combined), Turning left, Turning right} ∩ **DriveBehav1** ∉ {Improper backing, improper turn(combined)} ∩ **Traffic density** ∈ {Level A1, Unknown}, class: **crash.**
- Node 14: **PreInciManeuver** ∈ {Changing lanes, Negotiating a curve, Passing or overtaking another vehicle } ∩ **DriveBehav1** ∈ { Distracted, Exceeded speed limit, Other (combined), drowsy, exceeded safe speed, improper turn(combined)} ∩ **Traffic density** ∈ {Level A1, Level F, Unknown }, class: **crash.**
- Node 15: **PreInciManeuver** ∈ {Decelerating in traffic lane, Going straight (combined), Starting in traffic lane } ∩ **DriveBehav1** ∈ { Distracted, Exceeded speed limit, Other (combined), drowsy, exceeded safe speed, improper turn(combined)} ∩ **Traffic density** ∈ {Level A1, Level F, Unknown }, class: **Near crash**
- Node 16: **PreInciManeuver** ∈ {Entering a parking position} ∩ **DriveBehav1** ∈ {Distracted, Exceeded speed limit, Failed to signal, Following too closely, None, Other (combined), Right-of-way error, Stop sign violation, drowsy', exceeded safe speed, slowly-below speed limit} ∩ **Traffic density** ∈ {Level A2, Level B, Level C, Level D, Level E, Level F }, class: **crash**
- Node 17: **PreInciManeuver** ∈ {Other (combined), Turning left, Turning right} ∩ **DriveBehav1** ∈ {Distracted, Exceeded speed limit, Failed to signal, Following too closely, None, Other (combined), Right-of-way error, Stop sign violation, drowsy', exceeded safe speed, slowly-below speed limit} ∩ **Traffic density** ∈ {Level A2, Level B, Level C, Level D, Level E, Level F }, class: **Near crash**

Any path can be interpreted as a production rule which identify an interaction of human/external factors and driver behavioural mechanism yielding to a given response class. Joining all paths leading to the same response class label provides alternative scenarios causing a specific Event Severity.

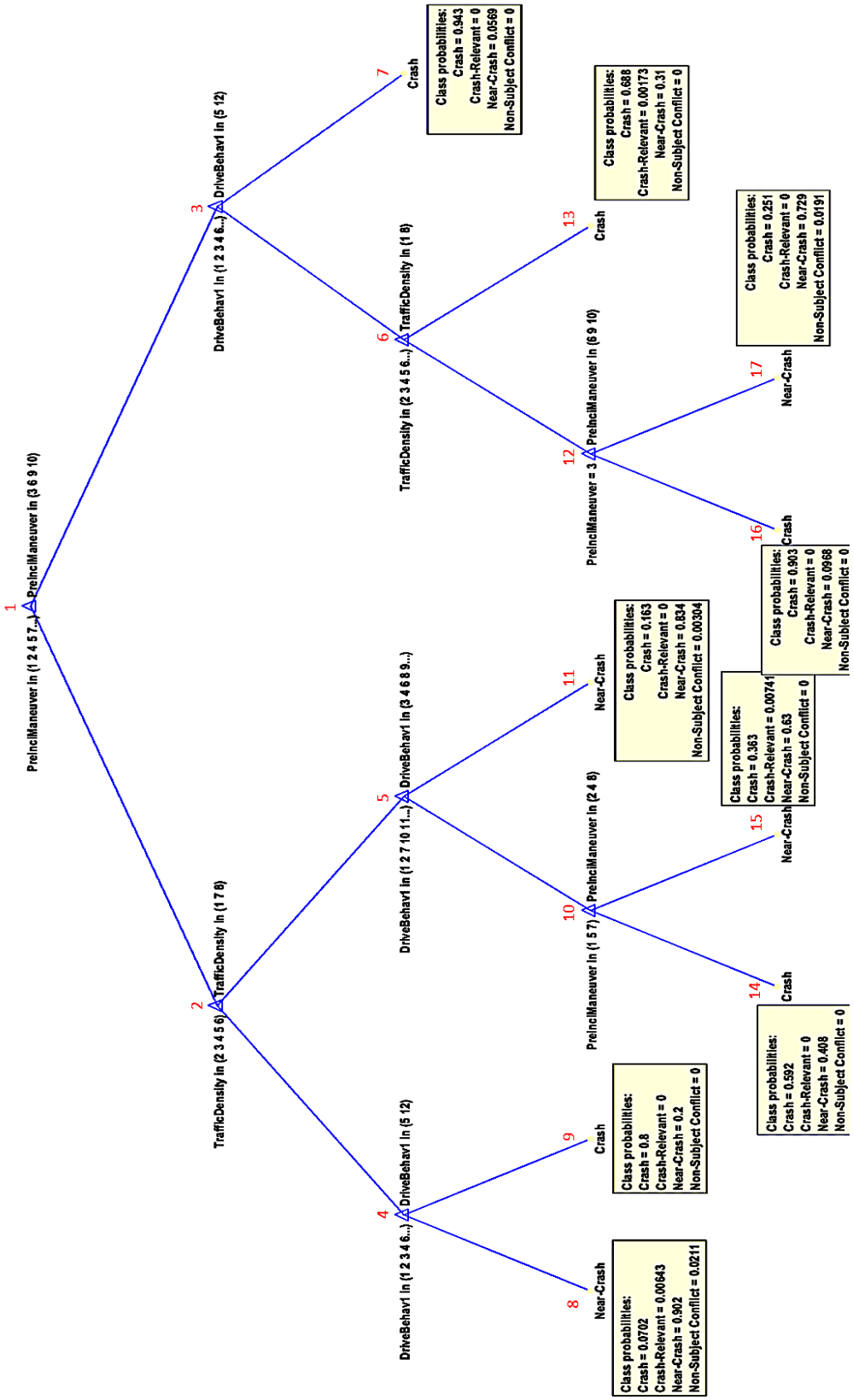

Figure 11. Classification Tree (Target Variable: Event Severity).

5.2.2 Random Forests

To assess the results of the classification tree, a Random Forests analysis has been performed on the same data, by setting the number of trees equal to 1000. Finally, the unbiased estimate of the importance of the variables was computed through permutations of out-of-bag predictor observations for random forests of regression trees (Loh, 2002).

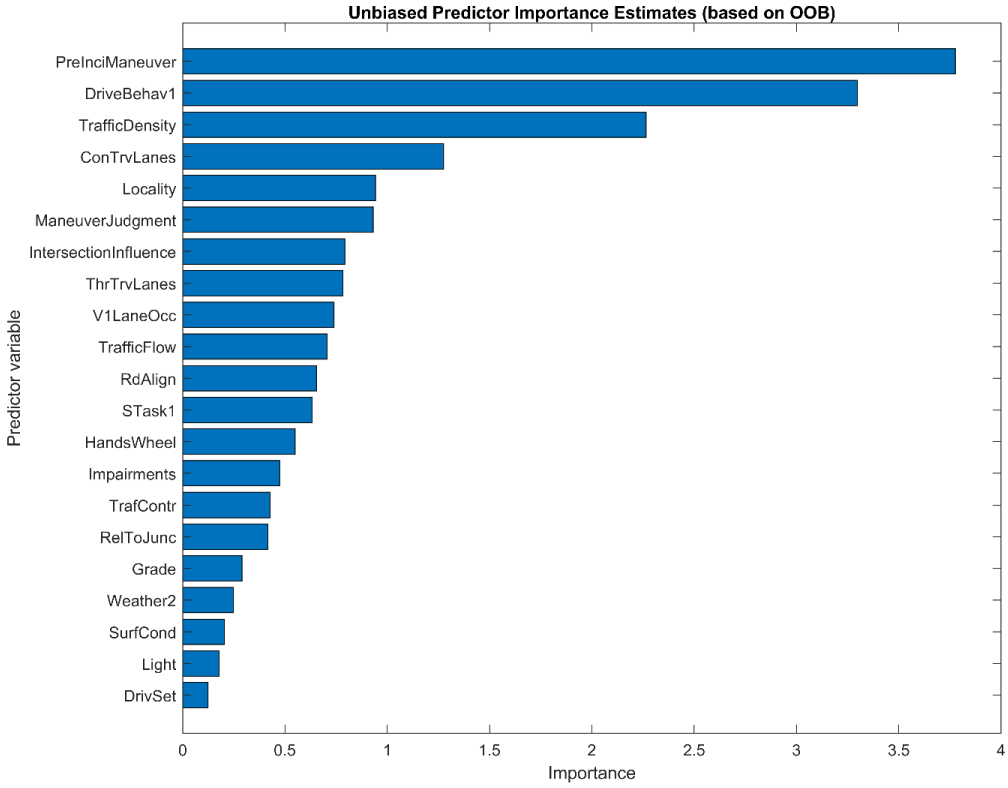

Figure 12. Random Forests Predictor Importance (Target Variable: Event Severity).

The importance of variables returned by the random forests let stronger the analysis made with the classification tree because the importance of the variables of the random forests confirms the robustness of the tree-based structure.

6. Next steps

More work on using global sensitivity analysis on machine learning is planned to use the SHRP2 Databases. The investigation will use both variance-based measures and the new discrepancy-based approach.

Works acknowledging i4Driving.

The project is acknowledged in:

a. two published papers

(Di Fiore et al. 2022) on sociology of quantification

(Saltelli et al. 2023) on impact assessment,

b. in two preprints:

(Saltelli and Puy 2022) on modelling, being revised for Humanities and Social Sciences Communication,

(Puy, Roy, and Saltelli 2023) about discrepancy measures for sensitivity analysis.

References

Aggarwal, C. (2015). Data Mining, Springer.

Aggarwal, C. (2018). Neural Networks and Deep Learning, Springer.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining Association Rules between sets of items in large databases, ACM SIGMOD, Vol. 22, 2, 207-216.

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2021). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety, 206, 107312.

Assi, K. (2020). Traffic crash severity prediction—A synergy by hybrid principal component analysis and machine learning models. *International journal of environmental research and public health*, *17*(20), 7598.

Bishop. C. (2007). Pattern Recognition and Machine Learning, Springer.

Becker, W., Paruolo, P., & Saltelli, A. (2021). Variable Selection in Regression Models Using Global Sensitivity Analysis. Journal of Time Series Econometrics, 13(2), 187-233.

Bénesse, C., Gamboa, F., Loubes, J. M., & Boissin, T. (2022). Fairness seen as global sensitivity analysis. Machine Learning, 1-28.

Borgonovo, E., & Iooss, B. (2016). Moment independent and reliability-based importance measures. In Handbook on Uncertainty Quantification (pp. 1-23). Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A. (1984). Classification and Regression Trees. Belmont CA: Wadsworth International Group.

Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis & Prevention*, *39*(4), 657-670.

Campolongo, F., Saltelli, A., & Cariboni, J. (2011). From screening to quantitative sensitivity analysis. A unified approach. Computer physics communications, 182(4), 978-988.

Chakraborty, P., Sharma, A., & Hegde, C. (2018, November). Freeway traffic incident detection from cameras: A semi-supervised learning approach. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1840-1845). IEEE.

Conversano, C., Mola, F., and Siciliano, R. (2001). Partitioning algorithms and combined model integration for data mining. *Computational Statistics*, 16(3), 323-339.

D'Ambrosio, A., Aria, M., & Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of Classification*, *29*(2), 227-258.

D'Ambrosio, A., Mazzeo, G., Iorio, C., and Siciliano, R. (2017a). A differential evolution algorithm for finding the median ranking under the Kemeny axiomatic approach. *Computers and Operations Research*, vol. 82, pp. 126-138.

D'Ambrosio, A., Aria, M., Iorio, C and Siciliano, R. (2017b). Regression trees for multivalued numerical response variables, *Expert systems with applications*, vol. 69, pp. 21-28.

D'Ambrosio, A., Iorio, C., Staiano, M., and Siciliano, R. (2019). Median constrained bucket order rank aggregation. *Computational Statistics*, pp.1-16.

Desrosières, Alain. 1998. The Politics of Large Numbers : A History of Statistical Reasoning. Harvard University Press.

Di Fiore, Monica, Marta Kuc Czarnecka, Samuele Lo Piano, Arnald Puy, and Andrea Saltelli. 2022. "The Challenge of Quantification: An Interdisciplinary Reading." Minerva 61 (December): 53–70. https://doi.org/10.1007/s11024-022-09481-w.

Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In Kittler, J. ands Roli, F (eds), International workshop on multiple classifier systems. Springer, Berlin, p. 1-15.

Efron, B. (1978). Computers and the Theory of Statistics: Thinking the Unthinkable, Stanford University. Department of Statistics.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans, SIAM.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.

Galante, F., Mauriello, F., Montella, A., Pernetti, M., Aria, M., & D'Ambrosio, A. (2010). Traffic calming along rural highways crossing small urban communities: Driving simulator experiment. *Accident Analysis & Prevention*, *42*(6), 1585-1594.

Gao, J., Yi, J., & Murphey, Y. L. (2023). Multi-scale space-time transformer for driving behavior detection. *Multimed. Tools Appl.*, 2023, 1–20.

Guo, F. (2019). Statistical Methods for Naturalistic Driving Studies. *Annu. Rev. Stat. Appl.*, 6(1), 309–328.

Guo, F. & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data, *Accid. Anal. Prev.*, 61, 3-9.

Hand, D.J. (1998). Statistics and More?, *The American Statistician*, 52,2,112-118, Taylor and Francis Group.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, New York: Springer.

Hoyt, C., & Owen, A. B. (2021). Efficient estimation of the ANOVA mean dimension, with an application to neural net classification. SIAM/ASA Journal on Uncertainty Quantification, 9(2), 708-730.

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety, 52(1), 1-17.

Iooss, B., Kenett, R., & Secchi, P. (2022). Different views of interpretability. In Interpretability for Industry 4.0: Statistical and Machine Learning Approaches (pp. 1-20). Cham: Springer International Publishing.

Iorio, C., Frasso, G., D'Ambrosio, A., and Siciliano R. (2016). Parsimonious Time Series Clustering using P-Splines, *Expert Systems with Applications*, vol. 52, pp. 26-38.

Iorio, C., Aria, M., D'Ambrosio, A., Siciliano, R. (2019). Informative Trees by Visual Pruning. *Expert systems with applications*, vol. 127, pp. 228-240.

Iorio, C., Frasso, G., D'Ambrosio, A., Siciliano, R. (2022). Boosted-oriented probabilistic smoothing-spline clustering of series, Statistical Methods & Applications, Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Statistical learning. In *An introduction to statistical learning* (pp. 15-57). Springer, New York, NY.

Lebart, §L., Morineau, A., Warwick, H. (1987). Multivariate Descriptive Statistical Analysis, Wiley & Sons.

Loh, W.Y. "Regression Trees with Unbiased Variable Selection and Interaction Detection." Statistica Sinica, Vol. 12, 2002, pp. 361–386.).

Lo Piano, S., Sheikholeslami, R., Puy, A., & Saltelli, A. (2022). Unpacking the modelling process via sensitivity auditing. Futures, 144, 103041.

Lo Piano, S., Puy, A., Sheikholeslami, R., & Saltelli, A. (2023). Sensitivity auditing: A practical checklist for auditing decision-relevant models. In A. Saltelli & M. Di Fiore (Eds.), The politics of modelling. Numbers between science and policy. Oxford University Press.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048-3061.

Mennicken, Andrea, and Robert Salais. 2022. The New Politics of Numbers: Utopia, Evidence and Democracy. Palgrave Macmillan.

Mola, F., and Siciliano, R. (1997). A fast splitting procedure for classification trees. *Statistics and Computing*, *7*(3), 209-216.

Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accident Analysis & Prevention*, *43*(4), 1451-1463.

Montella, A., de Oña, R., Mauriello, F., Riccardi, M. R., & Silvestro, G. (2020). A data mining approach to investigate patterns of powered two-wheeler crashes in Spain. *Accident Analysis & Prevention*, *134*, 105251.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention*, *49*, 58-72.

Montella, A., Aria, M., D'Ambrosio, A., Galante, F., Mauriello, F., & Pernetti, M. (2011a). Simulator evaluation of drivers' speed, deceleration and lateral position at rural intersections in relation to different perceptual cues. *Accident Analysis & Prevention*, *43*(6), 2072-2084.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2011b). Data-mining techniques for exploratory analysis of pedestrian crashes. *Transportation research record*, *2237*(1), 107-116.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2011). *Classification trees and association rules for exploratory analysis of powered two-wheeler crashes* (No. 11-0280).

Murphey, Y. L., Wang, K., Molnar, L. J., Eby, D. W., Giordani, B., Persad, C., & Stent, S. (2020). Development of Data Mining Methodologies to Advance Knowledge of Driver Behaviors in Naturalistic Driving. *SAE International Journal of Transportation Safety*, 8(2), 77–94.

Nisbet, R., Elder, J., Miner, G. (2009). Handbook of Statistical Analysis & Data Mining Applications, Elsevier.

Norton, J. (2015). An introduction to sensitivity assessment of simulation models. Environmental Modelling & Software, 69, 166-174.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. Science, 263(5147), 641-646.

Oreskes, N., & Conway, E. M. (2010). Defeating the merchants of doubt. Nature, 465(7299), 686-687.

Petrakos, G., Conversano, C., Farmakis, G., Mola, F., Siciliano, R., & Stavropoulos, P. (2004). New ways of specifying data edits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(2), 249-274.

Puy, A., Lo Piano, S., & Saltelli, A. (2020). Current models underestimate future irrigated areas. Geophysical Research Letters, 47(8), e2020GL087360.

Puy, A., Becker, W., Piano, S. L., & Saltelli, A. (2022). A comprehensive comparison of total-order estimators for global sensitivity analysis. International Journal for Uncertainty Quantification, 12(2).

Puy, A., Roy T.P & Saltelli, A. (2023). Discrepancy measures for sensitivity analysis, ArXiv https://arxiv.org/abs/2206.13470.

Ravetz, J., Funtowics, S. (1999). Post-Normal Science – an insight now maturing, *FUTURES,* vol. 31, pp. 641-646, Elsevier, open access.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., ... & Maier, H. R. (2021). The future of sensitivity analysis: An essential discipline for systems modeling and policy support. Environmental Modelling & Software, 137, 104954.

Rella Riccardi, M., Mauriello, F., Scarano, A., & Montella, A. (2022). Analysis of contributory factors of fatal pedestrian crashes by mixed logit model and association rules. *International journal of injury control and safety promotion,* 1-15.

Rella Riccardi, M., Mauriello, F., Sarkar, S., Galante, F., Scarano, A., & Montella, A. (2022). Parametric and non-parametric analyses for pedestrian crash severity prediction in Great Britain. *Sustainability,* 14(6), 3188.

Rella Riccardi, M., Galante, F., Scarano, A., & Montella, A. (2022). Econometric and Machine Learning Methods to Identify Pedestrian Crash Patterns. *Sustainability*, 14(22), 15471.

Saisana, M., Saltelli, A. (2008). Expert Panel Opinion and Global Sensitivity Analysis for Composite Indicators. In: Graziani, F. (eds) Computational Methods in Transport: Verification and Validation. Lecture Notes in Computational Science and Engineering, vol 62. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77362-7_11 .

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., ... & Tarantola, S. (2008). Global sensitivity analysis: the primer. John Wiley & Sons.

Saltelli, A. (2020), Ethics of quantification or quantification of ethics?", *FUTURES,* vol 116, pp. Elsevier, Open access, available online.

Saltelli, A., & Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12), 1508-1517.

Saltelli, A., & Di Fiore, M. (2020). From sociology of quantification to ethics of quantification. Humanities and Social Sciences Communications, 7(1), 1–8.

Saltelli, A., & Di Fiore, M. (Eds.). (2023). The politics of modelling. Numbers between science and policy, Oxford: Oxford University Press.

Saltelli, A., Kuc-Czarnecka, M., Piano, S. L., Lőrincz, M. J., Olczyk, M., Puy, A., ... & van Der Sluijs, J. P. (2023). Impact assessment culture in the European Union. Time for something new?. Environmental Science & Policy, 142, 99-111.

Siciliano R. and D'Ambrosio A. (2012). Statistical monitoring of tourism in the knowledge era. *In Morvillo A. (Ed.). Advances in Tourism Studies.* McGrow-Hill, pp. 231-258.

Siciliano, R., and Mola, F. (2000). Multivariate data analysis and modeling through classification and regression trees. *Computational Statistics & Data Analysis*, 32(3), 285-301.

Siciliano, R., D'Ambrosio, A., Aria, M. and Amodio, S. (2016) Analysis of web visit histories, part I: Distance-based visualization of sequence rules. *Journal of Classification*, vol. 33(2), pp. 298-324.

Saltelli, A., Giampietro, M., Gomiero, T., Forcing consensus is bad for science and society, The Conversation, May 11, 2017.

Sobol, I. M. (1993). Sensitivity analysis for non-linear mathematical models. Math. Modeling Comput. Exp., 1, 407-414.

Stark, P. B. (2018). Before reproducibility must come preproducibility. *Nature*, 557, 613–613.

Taleb, N. N. (2012). Antifragile: how to live in a world we don't understand (Vol. 3). London: Allen Lane.

Tukey, J.W. (1977). Exploratory Data Analysis, Pearson College Div.

Tunkiel, A. T., Sui, D., & Wiktorski, T. (2020). Data-driven sensitivity analysis of complex machine learning models: A case study of directional drilling. Journal of Petroleum Science and Engineering, 195, 107630.

Vapnik, V.N. (1995), The Nature of Statistical Learning Theory, New York: Springer Verlag.

Vapnik, V.N. (1998), Statistical Learning Theory, New York: Wiley.

Vokrinek, J., Schaefer, M., & Pinotti, D. (2014, May). Multi-agent traffic simulation for human-in-the-loop cooperative drive systems testing. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 1691-1692).

Wei, W., Larrey-Lassalle, P., Faure, T., Dumoulin, N., Roux, P., & Mathias, J. D. (2015). How to conduct a proper sensitivity analysis in life cycle assessment: taking into account correlations within LCI data and interactions within the LCA calculation model. Environmental science & technology, 49(1), 377-385.

Wu, Y., Zhang, J., Li, W., Liu, Y., Li, C., Tang, B., & Guo, G. (2023). Towards Human-Vehicle Interaction: Driving Risk Analysis Under Different Driver Vigilance States and Driving Risk Detection Method. *Automot. Innov.*, 6(1), 32–47.

Yang, J., Han, S., & Chen, Y. (2023). Prediction of Traffic Accident Severity Based on Random Forest. J. Adv. Transp., 2023, 1-8.