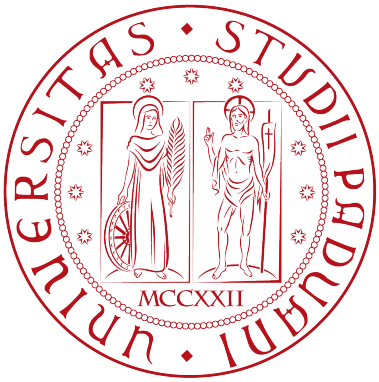


UNIVERSITÀ DEGLI STUDI DI PADOVA

SCUOLA DI INGEGNERIA
Corso di Laurea Magistrale in Ingegneria dell'Automazione

Data completeness assessment for real-world traffic scenarios

Academic Supervisor:
Prof. Schenato, Luca

Second Academic Supervisor:
Prof. Pillonetto, Gianluigi

Supervisor:
MSc. De Gelder, Erwin

Student:
Celin, Stefano

Matricola: 1180331

14 October, 2019

ACADEMIC YEAR: 2018-2019

Acknowledgement

Five years have already passed by since that day I got my first university ID number and a lot of persons have crossed their path with mine. I want to thank you all. You have made my academic years simply amazing.

Beside that, a special thank is needed to my colleague and mentor, Erwin, for all support, patient and for both personal and working opinions exchange. I am really glad I had the chance to work beside you.

I would like to thank Prof. Schenato Luca and Prof. Pillonetto Gianluigi for their feedback and time.

Last but definitely not least, I have to thank my family. You have made who I am today, and have showed me the path for a prosperous and successful tomorrow. A special thank to my mother, who would be really proud of my today's achievements and already looking forward to the incoming ones. I know you will be always here, beside me, enlightening me the right route.

Abstract

This thesis took part in collaboration with TNO - innovation for life in the Automotive Campus in Helmond, The Netherldands, as final project of the master in Automation Engineering, held at University of Padova, Italy.

Within the StreetWise project, data is used to develop tests and assessments helpful to claim the vehicle safety. Insufficient data may lead to misleading safety vehicle claims, therefore a representative measure of the database is needed. The data collection's aim is to enhance and improve nowadays safety assessment and reduce the fatalities mainly caused by the human presence within the control loop of the vehicles.

The StreetWise project's main goal is to collect all possible scenario that an Automated Vehicle may encounter while performing in everyday traffic situations.

As member of the StreetWise project, we analyse the completeness problem and define the used scenario-based approach. Two methods are proposed and only three assumptions are needed to compute the completeness measurement.

The first method requires a broaden overview of the project while the second one involves a statistical approach. Prior knowledge coming from experts and estimators used in biology were used to compute the number of possible scenario classes that an Automated Vehicle may have to cope with.

Both methods computed their levels of completeness regarding a database whose data concerns real-world traffic scenarios, helping future decisions on whether further data collecting is needed or not.

Contents

List of Figures

List of Tables

1 Introduction

Motivation and application

Mobility is highly correlated with societal and individual well-being and greatly contributes to equality of life. It is the backbone of commercial trading and services and, therefore, the basis for economic success [1]. Its advantages, however, come at price. As U.S. Department of Transportation [2] states, pedestrian deaths are the most since 1990, reaching, in the United States of America, 5987 in 2016 [2]. It has been estimated that 3258 motor vehicle occupants and motorcyclists who died in crashes in 2016 might have lived if they had used seat belts or motorcycle helmets [2].

The capability to overcome the human control interventions in critical situations, especially when the drivers are not fully aware of their physical and mental conditions, is becoming one of the goals of self-driving vehicles. The Automated Vehicle (AV) is no longer a dream and beside that, as the statistic says [2], is becoming every day more a need for driver safety rather than an additional car functionality.

According to 2015's predictions, in few decades, most cars are expected to be fully autonomous [3]. As pointed out in SAE J3016 [4] , AVs can be split into five different levels: driver assistance, partly automated driving, highly automated driving, fully automated driving and full automation [4]. The nowadays research focuses on completing the third level and is already looking at the next fully automated driving level. Nonetheless, what hinders the deployment of the fourth level and is raising concerns in researchers and final users are the safety vehicle claims that need to be fulfilled.

As Gelder *et al.* [5] mentioned, the amount of collected field data from driving studies is increasing rapidly and these data are extensively used for the research, development, assessment, and evaluation of driving-related topics; for example, see Ploeg *et al.* [6], Elrofai *et al.* [7], Gelder and Paardekooper [8], and Dingus *et al.* [9]. As mentioned by Alvarez *et al.* [10] and Geyer *et al.* [11], every work that is based on data, requires a measurement of the degree of completeness, especially when defining safety claims. When testing scenarios, based on collected data, the knowledge about the degree of completeness is requested and consequently questions like "when is the collected data enough?" are largely relevant.

The ongoing research topic addresses the previous question from an engineering point of view as *the data completeness problem*, which is also called *representativeness*, by finding a way to gauge the degree of completeness regarding a database in the automotive and Autonomous Driving (AD) field.

State of the art and challenges

Insufficient data may lead to inaccurate models, whereas excessive data lead to waste resources. Therefore, it is crucial to determine the right amount of data needed [12]. Achieving a representative database is attracting interest from various research environment such as medical [12]–[16] and biology [17]–[22] where researchers wonder how many different living animal species on earth there are.

Concurrently with the arrival of AVs and AD, the need of a completeness level started to rise concerns in traffic and transport environment. Kalra and Paddock [23] tried to claim the vehicle safety by using the number of miles that an AV has driven without accidents. It demonstrates that it is possible to drive hundred thousands of miles without experiencing new valuable driving situations. Studies regarding the collection of Naturalistic Driving Data (NDD) have been carried out [24] in the automotive field, with the aim to develop a behavioural driver model but the reply to:"how to measure the completeness of a database" in the automotive field still needs to be addressed. The stopping criteria proposed by Hauer *et al.* [25] used prior knowledge regarding the scenario distribution and probability to encounter new scenario that are not considered as assumption in this thesis.

Methodologies that consider a bigger picture of the problem, i.e. the prior knowledge that stay behind the data collection and a statistical approach that does not make use of the assumptions in [25], are missing.

Original contribution and goals

Two new approaches are carried out as possible solutions to compute the completeness level of a database. Using a broaden overview of the needed data completeness level in projects where data is used to define test/assessments and consequently claim the vehicle safety, the first method discloses a new methodology to exploit prior knowledge available to define how the completeness problem can be faced starting from the data collection to the computation of the degree of completeness. Through a case study, whose database concerns real-world traffic data, all steps needed to reach the final completeness level are illustrated.

Without requiring prior knowledge regarding the scenario distribution, the second method will cope with the completeness problem by using a statistical approach. Three well known estimators in biology, proposed by Chao and Lee [26] and Chao and Yang [27], are employed. The method exploits their estimators, used to evaluate the number of different animal species, to eventually compute the completeness level of the database. Unlike the first method, the second method does not require as much prior knowledge. Two case studies regarding the second method are carried out. The first one uses a subset of the whole real-world traffic database such that all estimators' parameters behaviour are

analysed when the completeness ground truth is known. The second case study involves the entire database and conclusions regarding the completeness level are drawn through a comparison with the previous case study.

Thesis outline

The chapters are organized as follows. Section 2 explains where the thesis is involved, and, in particular, in which project it belongs to. All relevant information for a thorough comprehension of the project are explained, definitions and methodologies which stand behind the scenario-based approach are illustrated. Finally, the completeness problem is formalized.

Section 3 firstly gives an introduction about differences amongst the two methods, respectively called *Strong Prior Knowledge* (SPK) and *Weak Prior Knowledge* (WPK) as well as the explanation of how each method copes with the completeness problem. Both the methods are explained in this section from a general perspective with few examples to strengthen the concepts.

As a proof of concept, both methods are applied to real-world traffic data collected and their respective completeness level is computed in Section 4. The same section highlights all important properties of the second method (WPK) by using part of the whole database. Thus, the properties can be recognised in the successive case, where the whole database is involved. Eventually, a discussion about the results and an overall final view of the two methods concludes the section.

Conclusions and future improvements regarding both the methods are drawn in Section 5.

2 StreetWise and Problem Formulation

Formally named $TNO\ -\ innovation\ for\ life$, is the largest Research & Technology Organization in the Netherlands and one of the largest in Europe. It connects people and knowledge to create innovations that boost the sustainable competitive strength of industry and the well-being of society.

TNO is recognized by Original Equipment Manufacturers all over the world as a valuable knowledge partner with unique expertise, tools and facilities. The activities within the research department Integrated Vehicle Safety (IVS) focus on improving the functional safety of vehicles and prevent human fatalities and severe injuries by developing integrated safety solutions that are robust and reliable under all circumstances.

In line with TNO's ambition to bring powerful innovation to the market, so does the StreetWise Project.

Nowadays, technologies are pushing every day further the human control role from the driving loop and, thanks to the transition towards the Connected and Automated Driving System, that moment is coming even faster [7].

The fundamental pillar of the StreetWise project is its new way to approach the problem. You could drive billions of kilometres and never encounter critical situations that the AVs should be able to deal with [23], therefore there is an essential need for constructing and collecting relevant traffic events and situations, called scenario, for testing and validating of AD functionalities [7].

A scenario is a quantitative description of the ego vehicle, its activities and/or goals, its dynamic environment (consisting of traffic environment and conditions) and its static environment. From the perspective of the ego vehicle, a scenario contains all relevant events [7]. Within the StreetWise project the scenarios are called 'real-world scenario' because they are extracted from real-world traffic data situations, i.e., data collected on the level of individual vehicles. They may include the road layout, weather and light conditions.

Gradually, with hours of driving data, the StreetWise database is becoming more and more representative for the situations that a vehicle may encounter while it is driving on public roads. Those data are eventually used for defining future tests and specific assessments for AD functionalities. Nonetheless, the database growth leads to other challenges like the need of automated mining and classification to process all data, as well as, the need of representativeness of the real world. The database should be able to describe as much as possible all the important and relevant scenarios that may happen along a common driver experience. It is therefore important to give a measure of completeness of the database: a metric that indicates how many of all possible scenarios the database

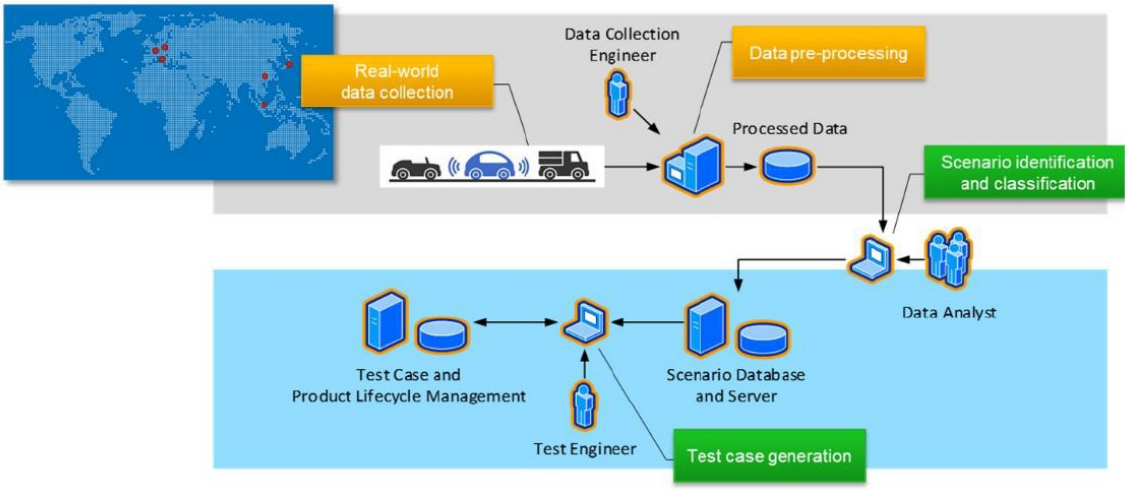


Figure 1: StreetWise pipeline: from the data collection to the test cases generation, the whole data path in the StreetWise Project [7].

already contains. The latter is the goal of this thesis.

Throughout this project, a description of the StreetWise methodology developed by TNO will be given, as well as how to create a scenario database from real-world data.

The data are collected by a fleet of vehicles in different cities, regions, and countries and are stored in a structured database for scenario analysis. As Figure 1 shows, the pipeline starts from real-world data, from which the data are collected, continuing with Data Analysis and eventually ending into TestCase and Product Lifecycle Management as last step.

From real-world data, collected from individual vehicles, scenarios are extracted.

2.1 Scenario

The scenario subjects are, first of all, the vehicles, which are grouped in two parts:

- Ego vehicle: It is the main character of the scene. It is the point of view from which the outside world is perceived through the vehicle sensors.
- Target vehicle(s): Other vehicles that surround the ego vehicle. They are part of the scenario because they are relevant in accordance with the ego vehicle perspective.

Secondly, scenarios include activities and events. An activity is defined as follow:

An activity is considered as the smallest building block of the dynamic part of the scenario (manoeuvre of the ego vehicle and the dynamic environment). An activity is a time evolution of state variables such as speed and heading to describe for instance a lane change, or a braking-to-standstill [7].

The term event is defined as follow:

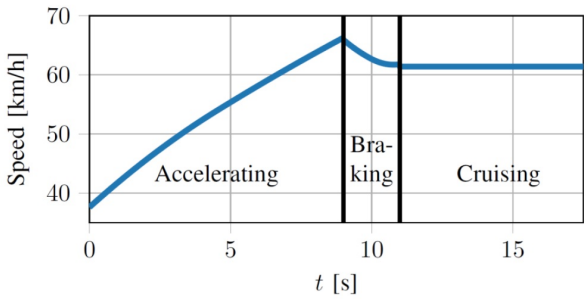


Figure 2: Speed profile example. According to the speed of the ego vehicle three activities can be extrapolated: accelerating, braking and cruising [7].

An event marks the time instant at which a transition of state occurs, such that before and after an event, the state corresponds to two different activities [7].

Remark 1. Examples of events are: the increasing speed for the acceleration, a rapidly decrease of speed for the braking and a constant speed regarding the cruising activity. ◊

Figure 2 illustrates an example of different activities delimited by events. In particular, the picture shows three different activities which are: accelerating, braking and cruising, respectively, bounded by events.

From a more practical aspect, the scenario is a time window in which particular manoeuvre done by either the ego vehicle or one of the target vehicles on the road take place. Within that time window, all the manoeuvres and other circumstances as weather and trajectories of the other traffic participants that may interact with the ego vehicle, are gathered and part of the scenario itself. The entire drive on the road can be, therefore, described by several scenarios where they may overlap in time. For instance, the ego vehicle may be facing a "braking at the front scenario" while another target vehicle is executing a cut-in at the front manoeuvre with respect to the ego vehicle's point of view. In this situation the two scenarios are overlapping in a short period of time.

Figure 3 shows what the scenario is composed of.

The scenarios are collected in a quantitative description and eventually grouped together as they share common properties.

This abstract group is called *Scenario Class* and it is formally defined as:

A scenario class refers to multiple scenarios with a common characteristic [7]

By adopting a tree structures of tags, the recognition of particular characterized scenario group is easier and the Figure 4 shows an example of tree structure regarding the dynamic traffic behaviour.

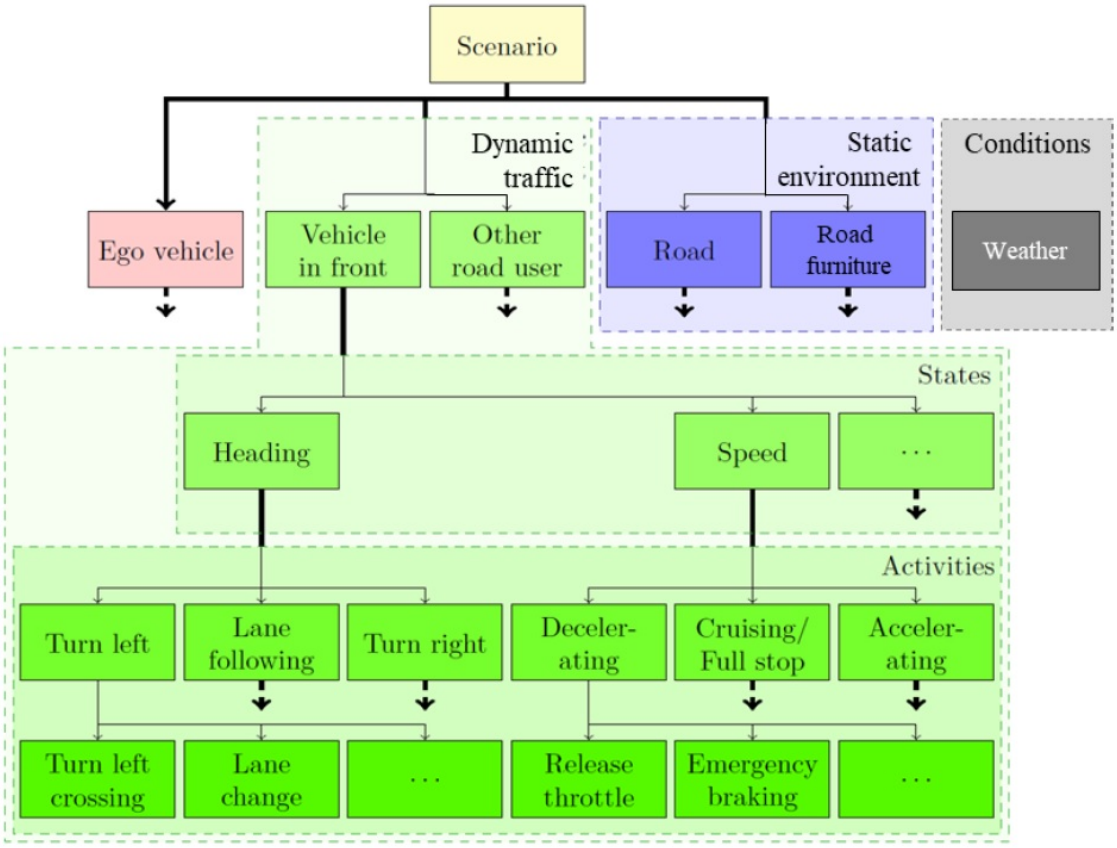


Figure 3: StreetWise pipeline [7].

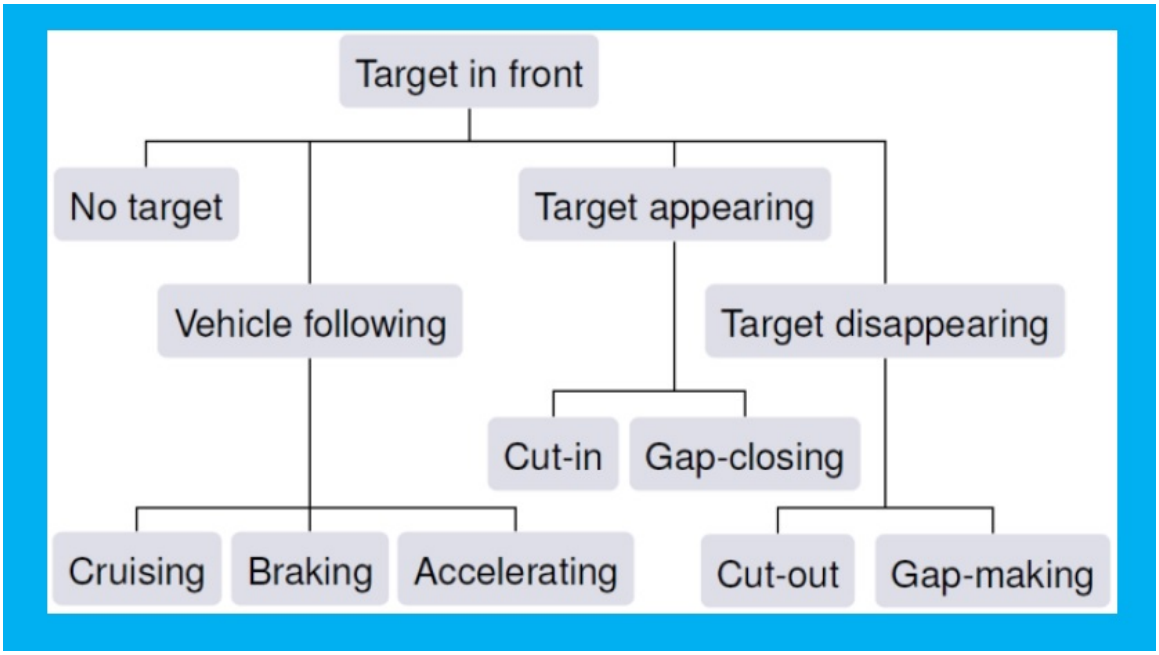


Figure 4: Dynamic traffic behavior. A tree structure in which only one option for each layer it can occur [7].

2.2 Data

The raw data that have been mentioned in the previous chapter are collected thanks to the fleet spread around the world, equipped with high-tech sensors. To describe the dynamic part of a scenario three main information are mandatory: accelerometer, camera, radar and Global Positioning System (GPS) [7]. Typically, other information are retrieved from the CAN-bus onboard, which manages detailed data such as individual wheel-speed, braking pressure and steering wheel rotation.

The raw data are not meant to be dealt in the StreetWise project, only object data are handled. The object data type includes crucial information for a faster data mining, such as an ID,a type (pedestrian,passenger car, truck, motor cycle, general objects, etc.), and state variable, i.e. relative position with respect to the ego vehicle, speed and heading of the object. All of them as a function of time.

Completing the scenario, sensors that acquire information about the static environment are necessary. The static environment is defined as follow:

The Static Environment is the line markers, lanes, road signal and, time-date stamp. Within this aspect it may be included every information which will not change during along the scenario. It may also include the presence of intersections, sidewalks, cycle lanes, tunnels, number of lanes etc [7].

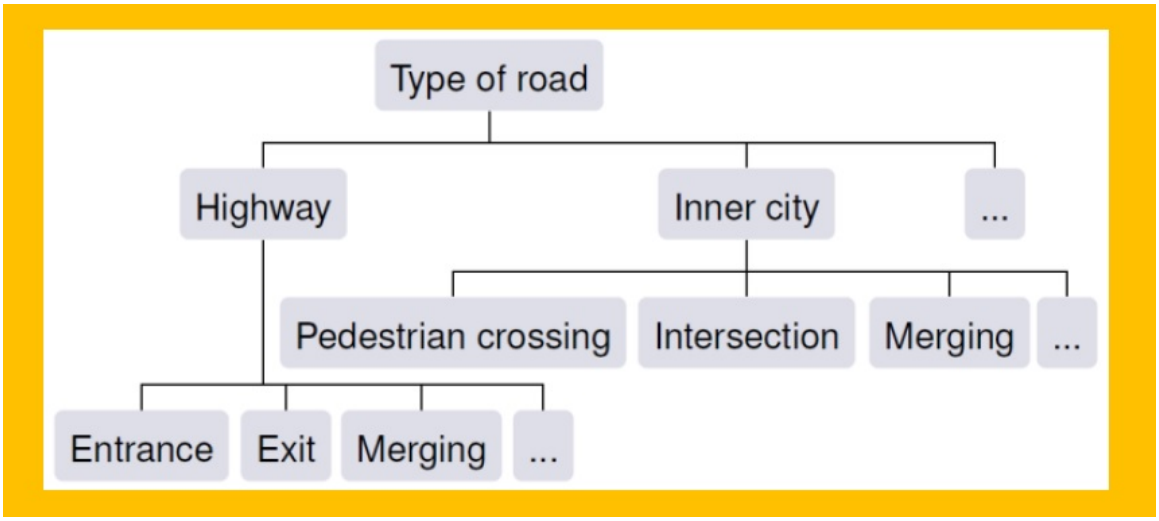


Figure 5: Static environment sketch based on a tree structure [7].

2.3 Scenario identification

How is it possible to retrieve activities and events from the gathered object data?

TNO has been developing techniques and algorithms to automatically detect events and activities in collected real-world traffic data. These are hybrid techniques that combine physical/deterministic models with data-analytics to detect events and activities hidden within terabytes of data.

The domain expertise of vehicle dynamics modelling as well as data analytics, machine learning and artificial intelligence, merged together, recently pointed out a new algorithm which is not only able to provide an overview of the type and frequency of events and/or activities, but also the parameters describing their characteristics. I.e. the maximum speed in lateral direction during a lane change indicates how aggressively the lane change is being performed. Parameters and activities are stored together for a easier a posterior data comprehension.

Figure 6 shows an overtaking being performed from a target vehicle. It highlights how the scenario mining works and example of activities and events are made clear. The overtaking takes place while entering in a tunnel thus the lightening conditions are changing so do the weather conditions.

Remark 2. Figure 6 is a good example of two scenarios overlapping each other. The first scenario can be defined by the ego vehicle which is performing a cruising on its lane, whereas the second scenario is the overtaking being performed by the target vehicle. ◊

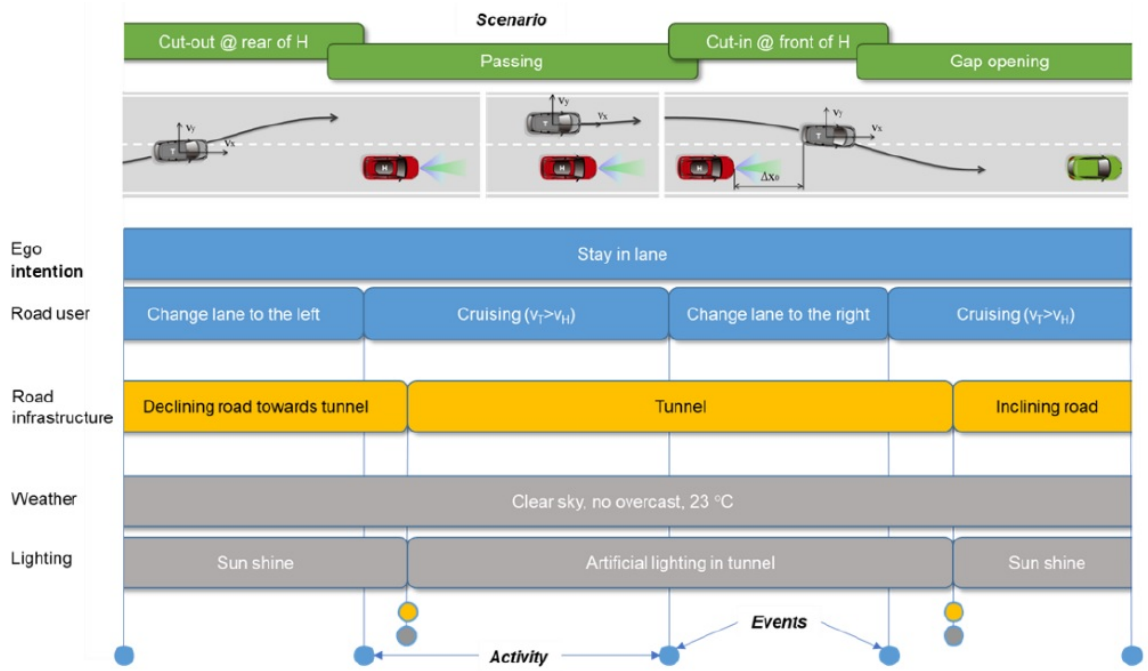


Figure 6: Scenario example. It highlights the events, the activities, and some different static environment circumstances like: entering the tunnel, inclining road [7].

2.4 StreetWise scenario database

The database contains real-world traffic scenarios from Europe, China, Japan, Singapore and US. The database provides a reliable view of scenarios that AVs may encounter on the road. More important are the differences in occurrence of these scenarios according to various cities, countries and continents that can be extrapolated from all those data. Amongst others, infrastructure layout, traffic rules, traffic behaviour and country specific conditions as well as climate or even culture, may affect the driver behaviour. To access all those information not only data with sufficient quality has to be retrieved but also the post-processing procedure needs to be able to parametrize all scenarios properly.

The StreetWise scenarios database does not contain raw sensor data, and some advantages of having object data instead of raw data are the following:

- The object data have only parametrized information about the scenario that they represent, therefore they are less sensitive.
- The object data use fewer parameters thus lower storage capability is required and easier is the post-processing computation.
- The parameters are time independent. They are, therefore, represented regardless of the original sample time.

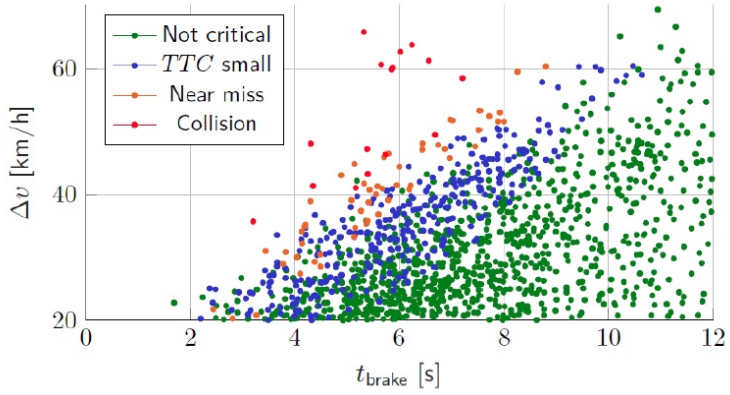


Figure 7: Parameter plot for test cases [7].

The fact that the scenarios are parametrized allows to other possibilities, indeed, the differences in scenario classes can be quantified. Surely, the database does not contain only critical or near-critical situations, therefore the most data will be about parametrization of every day behaviour on normal road. The "normal behaviour" parameters are extremely important as well as the critical occurrence parameters because they give support to the system developers about parameter ranges.

2.5 Test case generation

Dealing with high numbers and variety of scenarios may be misleading when test case generation has to take place. Some different methods in how to select among all the scenarios the one which fits best to the interesting case, are:

- Replay: a perfect replication of the same exact scenario, or even the whole test drive in order to achieve a model validation and or for assessment of a specific problem collected from a test drive.
- Extrapolation: exploiting the parameters which represent the scenarios regardless the sample time, by using statistical approach, it is possible to emphasize safety-critical cases or to select a limited amount of cases that are distributed over the complete operating area of an AD vehicle.
- Simulation: Similar as in Monte Carlo technique, a simulation model can cover the full ranges of the parameter distribution, highlighting relevant cases/areas such as critical cases for which the probability of a collision is high. The Figure 7 shows dots representing an example of the simulation results.

2.6 StreetWise project - Future goals

The incoming two years will be crucial for the StreetWise Project growth.

Using road types and operational design domains, TNO is building the automated algorithms to detect and classify the manoeuvres of the dynamic traffic participants. Starting from a highway road where only one direction is allowed and the lanes are well separated. The absence of crossing roads is crucial as well as no vulnerable road users are present. These conditions are no longer valid whereas we are dealing with urban roads where the manoeuvres can be almost unpredictable.

Other correlated projects can be merged with the StreetWise one in order to better achieve the urban environment modeling. One of them is the PROSPECT project [28] as well as automated valet parking applications are useful for making StreetWise fits for urban environments.

The relationship between multiple simultaneous actors in a scenario plays an important role. Nonetheless, in the StreetWise project the manoeuvres of those actors are treated as individual paths with implicit dependency which is not either described or modelled. It may be estimated though, and it is one of the next achievements of the project.

Approaching a totally different point of view the StreetWise project has been established dealing with individual scenario actors that receive information from the surrounding environment and infrastructure. The wireless communication brings another level of scenario generation in which information is being shared among different vehicles in order to enhance the vehicle's control as well as helping the driver during their daily manoeuvres.

2.7 Problem formulation

Roughly speaking the data completeness can be seen as either when all the data has been collected or when there is nothing left to learn. The latter occurs when, even though there might still be data that have not been gathered yet, those "unseen" data would not significantly change the completeness level. The completeness problem can be summarised, generically speaking, as: how much data is enough?How much more information can we retrieve while adding new data? These two questions, however, are quite wide and they have not been answered yet because of their complexity, see Section 1. The completeness problem is therefore split into three different subproblems [5]. According to StreetWise definitions the aforementioned subproblems are respectively concerning the completeness measurement regarding the activities, the scenarios and the scenario classes. One of them, the completeness problem regarding the activities, had already been addressed [5] and this thesis aims to answer the completeness problem regarding the scenario classes in the automotive field. Nonetheless, the problem formulation is formalised in its general perspective by referring to trace classes (regardless the field of application) instead of scenario classes.

The trace classes are a general group of classes that share common properties contained in the database D. The thesis solves the problem briefly summarised by the following question:

How to measure the data completeness with respect to the trace classes?

A general database D represents the starting point of this thesis work, while the ending point is a value, C, between 0 and 1, explaining the completeness level of the database concerning the trace classes. This value is obtained as outcome of the following equation:

$$C = \frac{S}{E}. \tag{1}$$

which is also called as OTC which stands for *observed trace classes* [29].

S is a subgroup of E representing the observed trace classes within the tests/assessments. Therefore, S needs to be always smaller or equal than E and it depends on D, $S(D)$. Different database leads to different subset S. E stands for how many different trace classes there are in total.

S is computed as a labelling process having the database D as input and the observed trace classes as output. E is computed according to the method that best suits each single situation and database.

On the one hand, choosing the statistical method, E considers statistical aspects and concepts since the *Weak Prior Knowledge* (WPK) method makes use of estimators. On the other hand, the *Strong Prior Knowledge* (SPK) method exploits all the available upfront knowledge to obtain E.

Remark 3. The completeness problem can be pictured as the dilemma concerning the estimation of animal species variety. D would be the database including each single capture, S would be the observed animal species and E would be the total number of animal species. $\Diamond$

2.7.1 Problem characterisation

The completeness problem deals with having a database D in which different information are stored and from which S (number of observed trace classes) is retrieved by defining a discrete model, ending up computing E (number of total trace classes) by using two different methods:

- Strong Prior Knowledge (SPK)
- Weak Prior Knowledge (WPK)

When both E and S are defined, the completeness level can be calculate as Eq. (1) explains.

Both methods exploit some assumptions that are made explicit as follows:

- Assumption 1: Given a log, the total number of all possible classes is finite
- Assumption 2: The event log is noise-free
- Assumption 3: Traces appear randomly and independently

Assumption 1: If the number of possible trace classes was infinite, given an even log, also called sample, (which always has a finite size), the ratio of observed trace classes to all possible trace classes would be always zero. It is therefore reasonable to assume that the number of possible trace classes is finite [29].

Assumption 2: Noise in an event log represents incorrect information, e.g. an event being missed or accidentally duplicated in the log. Noise is often caused by hardware failures and/or software bugs and the detection and remedy of these problems could lead one to also clean-up the event logs involved. It is reasonable to assume that the logs are noise-free as long as no general and feasible solutions for noise detection are released [29].

Assumption 3: It cannot be determined with absolute certainty what the next trace will be, based on already observed traces. It seems reasonable to assume that traces are produced randomly and independently.

Remark 4. S and E can be visualized as the number of different animal species captured versus the total number of living animal species. D represents the collected data concerning all captured animals. $\Diamond$

3 Data completeness: two methods

This chapter explains two methods whose prior knowledge, assumptions and methodologies are different. Figure 8 represents the main question whose reply determines which method is more suitable according to the test/assessment.

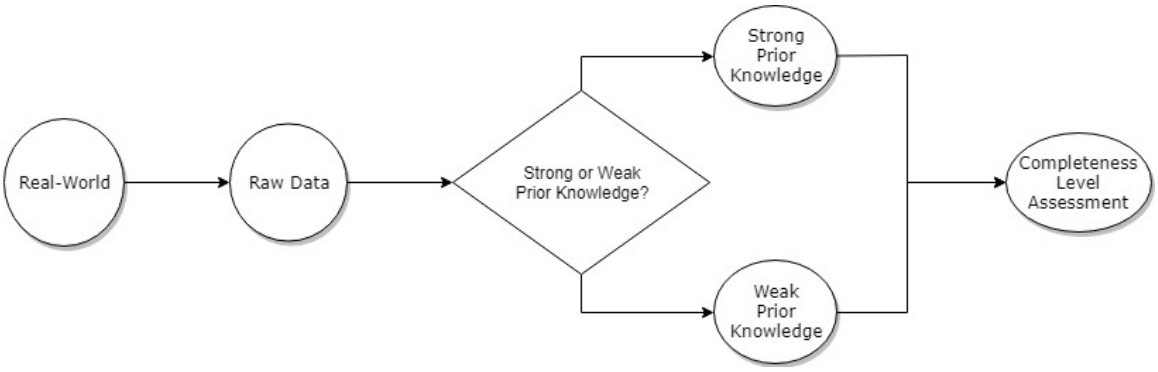


Figure 8: The main difference between the two methods: Strong or Weak Prior Knowledge.

As the name hints, the *Strong Prior Knowledge* foresees the use of a robust knowledge upfront, coming from the tests, assessments motivations and experts. Moreover, the method is not only suitable for databases concerning AVs and AD but also for a wide variety of databases whose completeness level is requested. Section 3.1 discloses each step from the raw data collection towards the computation of C where the prior knowledge, speaking for *the experts*, plays a decisive role in determining E. Not only the data accumulated during the test and experiments are important, all circumstances surrounding the experimentation produce the aforementioned prior knowledge. As Figure 9 illustrates, the only distinction between SPK and WPK methods concerns the usage of the expert knowledge.

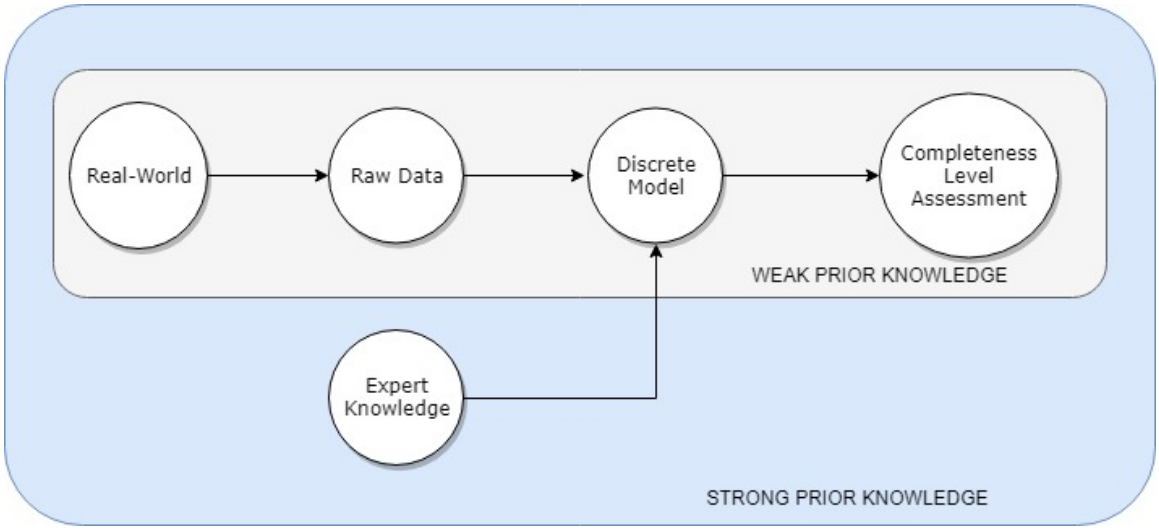


Figure 9: How the SPK and WPK deal with the completeness problem.

Unlike SPK, the second method WPK does not utilize the same prior knowledge and it involves estimators to obtain E. Furthermore, the SPK method gives a wide perspective and approach to the completeness problem while the WPK method is mainly related to estimate the total number of trace classes.

Remark 5. Continuing the animal species example, SPK foresees strong expertises regarding the captures and gathered animals, while WPK has solely D as knowledge. $\Diamond$

3.1 Strong Prior Knowledge

The SPK method is based on the claim that the completeness level evaluation only leans on motivations that make the data collection starting in the first place. The reason why the database has been created and the consequent tests/assessments uphold the whole method process and workflow.

Who are acquainted with all those motives will be called throughout this chapter and the following ones as *the experts*. The discrete model and the weighted function are delineated according to experts, which speak for prior knowledge, as the Figure 10 shows. Outputs of the discrete model are both S, the observed trace classes, that is strongly dependent on the database D, and E the total number of possible trace classes.

Eventually, before computing the degree of completeness, exploiting once more the experts knowledge, E can be split into L subgroups where each of those groups has its own importance. This process allows the enrich the completeness level when more important trace classes are observed by representing those level of importances through a weighted function.

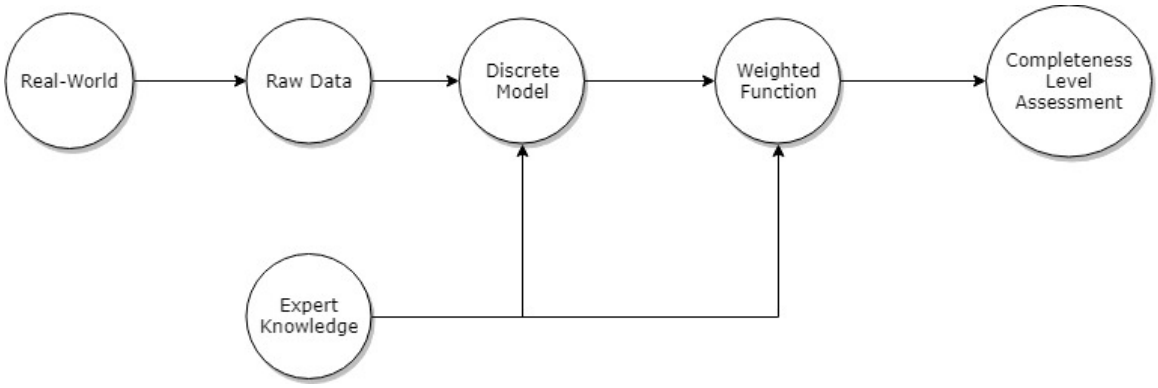


Figure 10: Strong prior knowledge method overview.

3.1.1 Prior knowledge and test boundaries

The motivations who led the experts starting the data collection and defining the assessment, regardless the data type, helps to delineate the discretization model process as well as the weighted function.

Clarifications about the requested prior knowledge by the SPK can be pointed out thanks to the following points:

- where the test takes place
- who is going to be the test lead
- what will be involved in the test
- what is the expected goal of the test
- what are all the circumstances surrounding the test
- what are the features being gauged along the test

If all the above points can be pointed out then the SPK method can be implemented and applied to D.

Remark 6. A prior knowledge regarding the animal species estimation example could be: geo-fenced area and only mammal animals. $\Diamond$

3.1.2 Model discretization

The need to create an abstract layer between the goal, the expert knowledge and the collected data is met through the discrete model. It allows to identify the trace classes observed and contained in the database D.

The distinction among classes allows to compute S since once that all irrelevant features are removed and the remained raw data have been mapped, S is consequently

known. Nonetheless, the distinction has to be unambiguous otherwise database items whose exact features are commonly shared with other ones may fall into different classes, eventually causing a different completeness level result [30].

The model discretization can be seen as a labelling process that has as input the database D and the observed trace classes, S, as output. It maps the raw data to its corresponding trace class.

E is computed exploiting the experts knowledge and according to the defined discrete model.

Remark 7. Defining the discrete model for the animal species estimation example foresees that, according to experts, information as whether conditions and other details, that do not help the trace class labelling process, should be discarded. Raw data as the length of the paws, the height and the colour of the animal can determine different animal species. Having as example of raw data: height higher than 2 meters and grey as colour, the discrete model would label it as elephant. By assuming only a discrete number of colours, heights and paws' length, the total number of possible animal species combinations is therefore computed, i.e., lions, hyenas, elephants and monkeys. $\lozenge$

3.1.3 Weighted function

The weighted function stands as the link between the prior knowledge and the assessment's goal. The chapter aims to properly define the weights such that the completeness level reflects as much as possible the entire project goals. In case this knowledge is not known upfront, the weights are chosen as uniform. If that is the case, Eq. (1) can be computed since both E and S have been established as Section 3.1.2 explained, obtaining the completeness level C of the database D.

This question the section wants to reply to states as follows:

- Among all the possible trace classes (E), are there any of them that have higher level of importance?

The weights are subject to the following equation:

$$\sum_{j=1}^{L} n_j w_j = 1 \tag{2}$$

L represents the total number of subgroups, n_j represents the number of classes that belong to the $j-th$ subgroup and w_j denotes the assigned weight to the $j-th$ subgroup.

The approach uses the Eq. (2), nonetheless, it requires additional prior knowledge from the experts to uniquely determine the weights.

Remark 8. Consider the example of estimating the number of animal species. Assuming that four-legged animal species are more important than two-legged ones, two macro

groups can be created and $L = 2$. w_1 is the weight associated to two-legged group, whereas w_2 is associated to four-legged one. Saying that lions (lionesses included),hyenas and elephants belong to the same four-legged group,$n_1 = 3$, whereas only monkeys belong to the two-legged group then $n_2 = 1$. Exploiting the experts knowledge and saying that $w_1 = 2 \cdot w_2$. The weights are $w_1 = 0.285$ and $w_2 = 0.143$. ◊

Remark 9. In case L is equal to E, n_j is equal to 1. A suitable choice for the weights w_j is a uniform and w_j are no longer j-dependent.

Then Eq. (2) becomes:

$$w \sum_{j=1}^{E} 1 = 1 \tag{3}$$

As expected, the weights are exactly:

$$w = \frac{1}{\sum_{j=1}^{E} 1} = \frac{1}{E} \tag{4}$$

◊

3.1.4 Completeness level

S and E have been defined in Section 3.1.2 and the weights in Section 3.1.3.

Defining $n_{s,j}$ as the number of observed trace classes in D that belong to the j-th group and w_j as explained in Section 3.1.3, the completeness level can be obtained as follow:

$$C = \sum_{j=1}^{S} n_{s,j} w_j. \tag{5}$$

Remark 10. In case the weights are all equal and correspond to $\frac{1}{E}$ (Eq. (4)) then $n_{s,j}$ would become 1 and consequently:

$$C = \sum_{j=1}^{S} n_{s,j} w = w \sum_{j=1}^{S} 1 = wS = \frac{S}{E}, \tag{6}$$

which is Eq. (1). ◊

Remark 11. Assuming that only hyenas and monkeys have been captured then $n_{s,1} = n_{s,2} = 1$. The overall completeness level would be $0.285 \cdot 1 + 0.143 \cdot 1 = 0.428$. If both hyenas and lions had been seen and no monkeys, then the completeness level would have been $0.285 \cdot 2 + 0.143 \cdot 0 = 0.57$. ◊

3.2 Weak Prior Knowledge

When the prior knowledge is either not available or the total number of trace classes cannot be computed directly from the discrete model, the SPK method is not suitable. Therefore, the WPK method, which uses a statistic approach, is used. This method can estimate that number E by only using the observed classes S and their occurrences without requiring different assumption but the ones stated in Section 2.7.

Determining the degree of completeness can be compared as the problem of guessing from an urn with an unknown but finite number of marbles, how many different colours these marbles may have and how many marbles there are of each colour. The only input to the guessing problem is a number of selections of marbles where it was recorded what the colour of the marble was before it was put back in the urn [31].

This marble-guessing problem constitutes a particular occurrence of the so-called *species animal estimation problem*, already used as example throughout Section 3.1, wherein the number of species of a population, based on a finite sample, needs to be estimated [29].

Even though this is a well-known problem in the field of statistics [19], [32] not always the same assumptions and goals are shared among similar topic. It needs to be noted that there is a difference between estimation of the number of animals and estimation of the number of animal species. The database reveals detailed information concerning the type of the problem. Since the analysed database D may contain twice the same trace class within the same draw then the correlated problem is similar to the animal species estimation one. If the database had only once the same trace class then it would be as the animal estimation problem.

Remark 12. In the same draw it is not possible to capture twice the same animal, however, it may happen that two animals of the same species are caught. ♢

As Section 3 previously pointed out, WPK does not foresee the use of any prior knowledge and the workflow can be represented as Figure 11 shows.

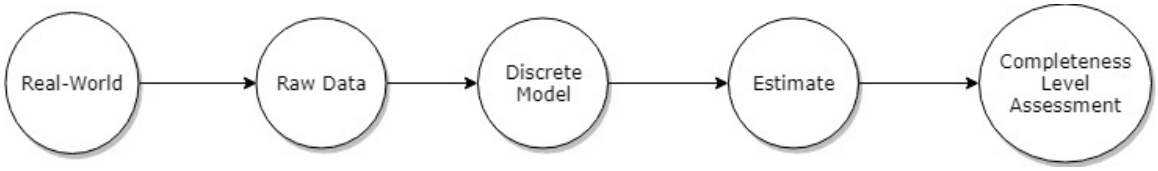


Figure 11: Weak prior knowledge method overview.

Figure 11 highlights that no expert knowledge is present and neither is the weighted function. An uniformly distributed function will be used.

The input of the whole method is the database D, whose degree of completeness is demanded.

3.2.1 Model discretization

Although the prior expertise is absent, a way to represent the data and to properly assign the database items to their classes is requested, so it is the discrete model. Moreover, the discretization procedure returns as output S, while E is obtained as the outcome of the estimation process.

The WPK's discrete model uses only the database D as prior knowledge.

The model discretization defines, as it was for the SPK method, the distinguishing line between classes as well as the labelling process from the raw data to their appropriate class.

Amongst all the collected features, only the relevant ones need to be kept and the other ones discharged. It is important to know that as in the SPK method, different discrete models lead to diverse completeness levels, even though the same database D is used.

3.2.2 Estimators

This section goes deeper in details concerning the different estimators for E. A wide overview of all possible estimators can be found in [29].

The characteristics of the completeness problem exclude some approaches used for similar estimating problems.

- since no priori occurrence distribution of trace classes is assumed for the problem of completeness, approaches depending heavily on the presence of specific occurrence distributions are not considered to be suitable. For example, approaches requiring that occurrence distributions of trace classes are known [33] and approaches based on parametric models [34].
- Approaches assuming sampling without replacement are ignored. Sampling on a finite population without replacement would mean that a marble sampled later has a different occurrence probability from the one of the same class previously captured [35].

Eventually, the estimators used to accomplish the completeness problem are the following ones:

- Chao and Lee's estimators [26]
- Chao and Yang's estimators [27]

3.2.3 Chao and Lee's estimators

In Chao and Lee's estimations [26], a random sample is drawn from a population with an unknown number of classes and unequal class probabilities. A nonparametric estimation technique is proposed to estimate the number of classes using the idea of sample coverage.

A random sample of size n is taken from a population whose number of classes is N. The classes are indexed by $1, 2, \ldots, N$; p_i represents the probability that any observation belongs to the i-th class with $i \in \{1, 2, \ldots, N\}$; X_i is the number of elements of the i-th class observed within the sample. f_i is the number of classes that have exactly i elements in the sample and it is calculated as:

$$f_i = \sum_{j=1}^{N} I[X_j = i],$$

where $I[A]$ is the indicator function [26] that returns 1 if A is true and 0 otherwise.

The number of distinct classes observed in the sample is:

$$F = \sum_{i=1}^{\infty} f_i,$$

and therefore the sample size can be written as:

$$n = \sum_{i=1}^{\infty} i f_i.$$

Note that even though the sum index goes to ∞, f_i is 0 when $i > n$.

The estimator's goal is to achieve the number of classes represented within the sample, N, based on f_i with $i \geq 0$. Once that N has been estimated it is used as E in the Eq. (1).

The equally likely or equiprobable assumption, i.e., when $p_i = p_j \; \forall \; i, j \in \{1, 2, \ldots, N\}$, has been discussed by many authors [22], [36] and it is used to compute, later on this section, the first estimator N_1.

More related to the completeness problem is instead when the equiprobable assumption is not valid.

C_s stands for the sample coverage and it is defined as:

$$C_s = \sum_{i=1}^{N} p_i I[X_i > 0].$$

The above highlights that C_s varies with the sample X_i and it is a random variable so C_s needs to be estimated. A method is made clear as follow:

$$\hat{C}_s = 1 - \frac{f_1}{n}. \tag{7}$$

As Chao and Lee [20] bears out, if we estimate the number of classes without estimating the variation among the class probabilities, we can derive only a lower bound, with the lower bound being achieved in the equiprobable case.

In the equiprobable case, $p_1 = p_2 = \ldots = p_N = 1$, C_s becomes:

$$C_s = \sum_{i=1}^{N} p_i I[X_i > 0] = \frac{1}{N} \sum_{i=1}^{N} I[X_i > 0] = \frac{1}{N} \sum_{i=1}^{\infty} f_i = \frac{F}{N}$$

therefore the first estimator of N is:

$$\hat{N}_1 = \frac{F}{\hat{C}_s}. \tag{8}$$

The estimator above represents the starting point from which the final estimator is defined as its enhancement.

Assuming that a random sample of size n is drawn from a population of N cells with fixed cell probabilities $\mathbf{p} = (p_1,p_2,...,p_N)$, $\sum p_i = 1$, letting p_1, p_2, ..., p_N have mean $\bar{p} = \sum_i p_i/N$ and Coefficient of Variation (CV) $\gamma = \frac{[\sum_i (p_i - \bar{p})^2/N]^{1/2}}{\bar{p}}$, Chao and Lee [26] prove that:

$$\mathsf{E}\left[\frac{\sum_{i=1}^{\infty} i(i-1)f_i}{(n(n-1))}\right] = \sum_{i=1}^{N} p_i^2,$$

therefore

$$\gamma^2 = N \sum p_i^2 - 1 = \frac{N \sum_{i=1}^{\infty} i(i-1)\mathsf{E}[f_i]}{[n(n-1)]} - 1 \tag{9}$$

Remark 13. $\gamma = 0$ means that all p_i's are equal. ◊

Remark 14. Note that $\mathsf{E}[\cdot]$ represents the expectation value of the quantity. ◊

N in Eq. (9) is not available so it is substituted by $\hat{N}_1$, ending up with the following estimators of the non-negative parameter γ^2:

$$\hat{\gamma}^2 = \max\left\{\frac{\hat{N}_1 \sum_{i=1}^{\infty} i(i-1)\mathsf{E}[f_i]}{[n(n-1)]} - 1, 0\right\} \tag{10}$$

$$\tilde{\gamma}^2 = \max\left\{\hat{\gamma}^2 \left(1 + \frac{n(1-\hat{C}_s)\sum_{i=1}^{\infty} i(i-1)f_i}{[n(n-1)\hat{C}_s]}\right), 0\right\} \tag{11}$$

When the true value of CV is relatively large, it is suggested to use $\tilde{\gamma}^2$ [26].

Eventually, the two estimators are the following ones:

$$\hat{N}_2 = \frac{F}{\hat{C}_s} + \frac{n(1-\hat{C}_s)}{\hat{C}_s}\hat{\gamma}^2. \tag{12}$$

$$\hat{N}_3 = \frac{F}{\hat{C}_s} + \frac{n(1-\hat{C}_s)}{\hat{C}_s}\tilde{\gamma}^2. \tag{13}$$

The variance of Eq. (12) and Eq. (13) can be computed by using an asymptotic approach. It is important to note that both the above estimators are a function of f_1,f_2,...,f_n, so, assuming γ positive, the following equations can be adopted:

$$\hat{N}_2 = \hat{N}_2(f_1, f_2, ..., f_n) \tag{14}$$

$$= \frac{\sum_{i=1}^{\infty} f_i}{1 - f_1/n} + \frac{f_1}{1 - f_1/n}\left(\frac{\sum_{i=1}^{\infty} f_i}{1 - f_1/n}\frac{\sum_{i=1}^{\infty} i(i-1)f_i}{n(n-1)} - 1\right). \tag{15}$$

$$\hat{N}_3 = \hat{N}_3(f_1, f_2, ..., f_n) = \tag{16}$$

$$= \frac{\sum_{i=1}^{\infty} f_i}{1 - f_1/n} + \frac{f_1}{1 - f_1/n}\left[\frac{\sum_{i=1}^{\infty} f_i}{1 - f_1/n}\frac{\sum_{i=1}^{\infty} i(i-1)f_i}{n(n-1)}\left(\frac{\sum_{i=1}^{\infty} i(i-1)f_i}{n(n-1)(1 - f_1/n)} + 1\right) - 2\right]. \tag{17}$$

The variance can be computed as follow:

$$\hat{\text{var}}(\hat{N}_l) \approx \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial \hat{N}_l}{\partial f_i}\frac{\partial \hat{N}_l}{\partial f_j}\text{cov}(f_i, f_j), l = \{2, 3\}, \tag{18}$$

and the covariance is computed as:

$$\text{cov}(f_i, f_j) = \begin{cases} f_i(1 - f_i/\hat{N}_2), & \text{if i = j} \\ -f_i f_j/\hat{N}_2, & \text{if i} \neq \text{j} \end{cases}. \tag{19}$$

f_1 plays an important role as Eq. (14) shows. If f_1 gets 0 then $\hat{N}_2$ and $\hat{N}_3$ become:

$$\hat{N}_2 = \sum_{i=1}^{\infty} f_i,$$
$$\hat{N}_3 = \sum_{i=1}^{\infty} f_i.$$

The two estimators therefore coincide, if $f_1 = 0$.

3.2.4 Chao and Yang's estimators

Chao and Lee [26] estimators use all occurrences f_i to estimate the number of classes N. Nonetheless, it must be noted that only the first subset of occurrences plays an important role in the estimation process, therefore an improvement of their estimators has been carried out by Chao and Yang [27] and this subchapter explains its functioning.

The Chao and Yang's estimator [27] separates the species into two groups according to their f_i. The groups are split so that there is the group whose f_i (number of occurrences) is lower than κ and the group whose f_i is higher than κ. This grouping rose up importance because, in order to detect the unobserved data, the higher value of the index i in f_i are ($f_i \neq 0$), the weaker is their contribution in the estimation process. As assumption it is known that some classes are more likely to happen than other ones and this is made clear through the f_i values. Determining the quantity of different classes, concerns the estimation of how many classes there still are that have not been observed yet, therefore the higher is f_i ($i \in \{1, \ldots, \kappa\}$) the higher are the odds to find new classes within the next draws.

Remark 15. As example, having f_1 close to zero means that not many classes are still unobserved, whereas, having $f_{10000} = 1000$ does not help any decision-making because it only shows that the more likely to occur, have occurred during the test as expected. $\Diamond$

This estimator emphasis low frequency species which are deemed to be important for estimating the number of unobserved species. Positively biased estimates are caused by not separating the long-tailed data [29].

As in Section 3.2.3, a random sample of size n is taken from a population whose number of classes is N. All previously definitions still hold, i.e. X_i, F and $I(\cdot)$.

The goal is to estimate the number of N, which is necessary to compute the completeness level as Eq. (1) shows.

The Chao and Yang's estimator [27] is an improvement of the Chao and Lee's one and the equiprobable case has the same estimator:

$$\hat{N}_1 = \frac{F}{C_s},$$

while C_s is the same as Eq. (7).

A suitable value for κ might be 10 [27] and, in Section 4, will better argue this choice. The number of classes detected more than κ times are added to the estimation. Having a subset of classes means that CV is smaller than the original one (used in Section 3.2.3).

The improvement, respect to the Section 3.2.3 arrives when γ is higher than 1. The used estimator is no longer Eq. (14) but instead:

$$F^* = F - \sum_{i=1}^{N} I(X_i > \kappa) = \sum_{i=1}^{N} I(0 < X_i \leqslant \kappa),$$

are the number of distinct classes detected within the subset.

$$\hat{C}_s^* = 1 - \frac{f_1}{\sum_{i=1}^{\kappa} i f_i},$$

is the sample coverage estimator.

The estimator for N becomes:

$$N_1^* = \sum_{i=1}^{N} I(X_i > \kappa) + \frac{F^*}{\hat{C}_s^*}, \tag{20}$$

and

$$N_2^* = \sum_{i=1}^{N} I(X_i > \kappa) + \frac{F^*}{\hat{C}_s^*} + \frac{f_1}{\hat{C}_s^*}\hat{\gamma}^{2^*}, \tag{21}$$

where

$$\hat{\gamma}^{2^*} = \max\left\{\frac{F^*}{\hat{C}_s^*}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{(\sum_{i=1}^{\kappa} i f_i)^2} - 1, 0\right\}. \tag{22}$$

Note a small difference between Eq. (22) and Eq. (10) in the denominator. Eq. (22) presents an approximation that leads to a smaller estimation since γ is lower.

The variance of both the estimators can be computed as Eq. (18) pointed out.

The difference between Chao and Lee's estimators [26] and Chao and Yang's [27] ones is expressed by κ because it splits the entire database into two subsets. Saying that the first occurrences are the ones that tell more about the unobserved classes, Chao and Yang exploit this as enhancement of the Chao and Lee's estimator [26].

The closer κ gets to N the similar are the behaviours of the two estimators. Substituting κ with N in Eq. (20) both methods are the same because $I(X_i > N) = 0$ and therefore $F^* = F$.

3.2.5 Completeness

E has been computed through Section 3.2.2, whereas S has been obtained via Section 3.1.2. No weighted function is applied and therefore each class has the same importance than the other ones. Calling back Section 3.1.3 the used equation is Eq. (3).

Important to note, when the ground truth is unknown, might be that higher estimations lead to lower completeness levels, therefore underestimations may happen. Nonetheless, an underestimation might be an advantage that is preferable over an overestimation be-

cause it is more conservative and it may lead the conclusion to gather more data. On the other case, overestimation may lead to stop the data collection and consequently, affect future decisions.

4 Case study: real-world traffic scenarios

This section applies each aforementioned method in the automotive field context. In particular, the database whose completeness is computed as case study, concerns data regarding AVs and AD data, gathered as part of the StreetWise project. The trace classes are defined according to Section 2 as scenario classes and a discrete model example is developed and considered. Eventually, the obtained completeness levels and differences among the estimators are illustrated.

4.1 Strong Prior Knowledge

The Strong Prior Knowledge method exploits all expertises known upfront. Section 4.1.1 describes the database D in details, especially the limitations and circumstances that are useful to determine the discrete model (Section 4.1.2) and the weighted function (Section 4.1.3).

4.1.1 The database and test boundaries

The database D, used as Proof-of-Concept (PoC), took part as particular project in the StreetWise context, whose minor aim was to develop a Personalized Adaptive Cruise Control (PACC) according to the driver preferences [37]. The major purpose was to enlarge the TNO's database, enriching the knowledge of observed scenarios.

Twenty different drivers between 25 and 60 years old, having their driving license for more than 5 years and driving more than 5000 km a year, were involved in the data collection. They drove approximately half in manual mode and half in Adaptive Cruise Control (ACC) mode and the test was performed during daytime (8.30am-5pm) under dry weather conditions.

The vehicle was a dedicated car equipped with high-tech sensors that allowed to gather data as the following ones, creating the database D.

- velocity
- accelerations
- yaw rate
- steering wheel angle
- wheel speeds
- object/lane information

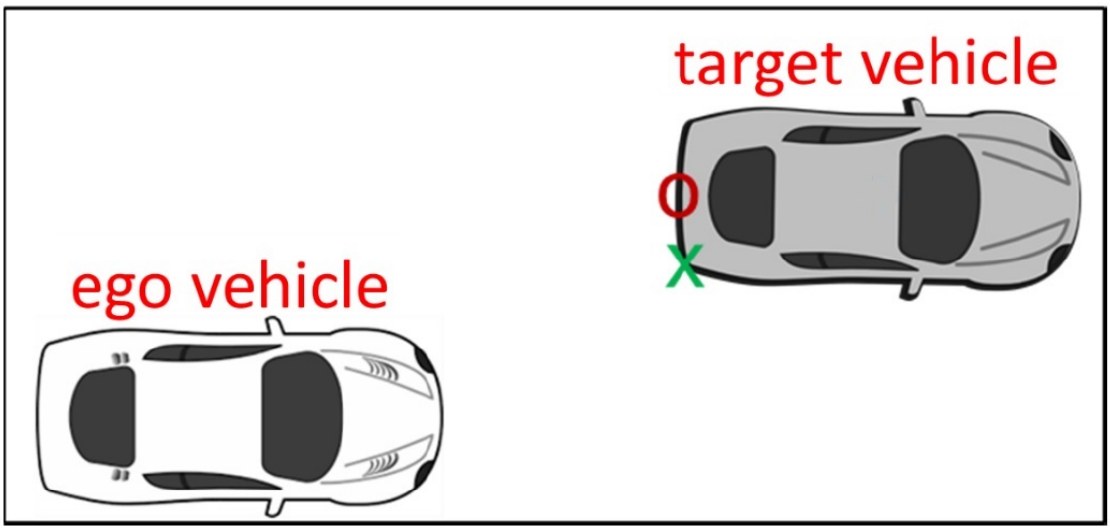


Figure 12: How the relative distances between cars are defined. The red circle as reference point for longitudinal distance while the green cross is used for the lateral one.

- LIDAR
- GPS
- radar: centre, rear left, rear right.

Beside the above, as prior knowledge, it is known that the drivers drove a fixed route in the region of Amersfoort in the Netherlands with a length of 46km and about 50 minutes of driving without traffic issues. 55% was highway, 40% was on rural roads and the remaining 5% was in urban areas. 55% of the route consists of multi-lane unidirectional roads.

Remark 16. A connection with Section 2 is made by associating all above information into the correct definition, i.e. weather and lighting conditions as *conditions*, road layout and static elements as *static environment* and, manoeuvres of actors and traffic situations as *dynamic environment*. ◊

The longitudinal distance is considered as the distance between the target vehicle's rear bumper centre and the ego vehicle's front bumper centre. Similar is detected the relative distance when a target vehicle is approaching the ego vehicle from behind. For the lateral position, the bumper later edge of the vehicle are used as reference points. Figure 12 shows the aforementioned reference points.

All above represents a prior knowledge useful to apply the SPK method on the database D and, beside the PACC, the main goal of the StreetWise project, as briefly introduced in the section preface, concerns the vehicle safety. As Figure 13 shows, the completeness level represents the feedback from the vehicle safety claim to the test and, furthermore,

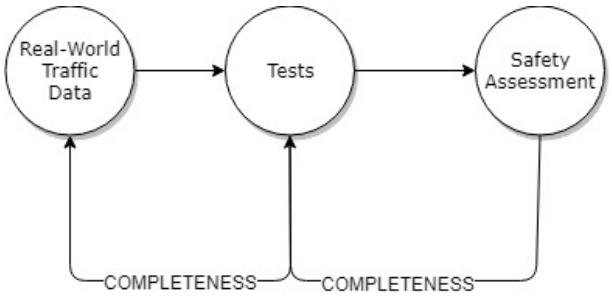


Figure 13: How the completeness affects the overall StreetWise Project.

from the test to the real-world traffic data. If the completeness is not achieved not all representative tests can be developed and therefore a complete vehicle safety can not be guaranteed.

4.1.2 Model discretization

To properly solve the completeness problem, briefly summarised as:"*how to measure the completeness regarding the scenario classes*" it is necessary to define the discrete model as Section 3.1.2 explained.

The discrete model has to label the raw data contained in the database D thus the observed scenario classes, S, can be obtained.

A discrete model solution is represented in Figure 14 and it is called *scene*. It represents the ego vehicle in the centre and the discrete model: three lane-road layout as vertical discretization and three dot blue lines as longitudinal discretization.

Remark 17. The discrete model is exploiting the fact that more than 50% of the collected data in D concerns about highway roads, so three lanes are drawn. $\lozenge$

The scene splits into twelve parts the road surrounding the ego vehicle and, letting d denote the relative distance, as Section 4.1.1 explained, a possible choice for the horizontal discrete thresholds follows:

$$7 \leq d \leq 15 \tag{23}$$

$$0 \leq d < 7 \tag{24}$$

$$-7 \leq d < 0 \tag{25}$$

$$-15 \leq d < -7 \tag{26}$$

The positive value indicates the space at the front of the ego vehicle and the negative the part behind. The distance d has been chosen so that approximately only one car can fit in each part and the choice of including the extreme values does not affect significantly,

therefore, it is arbitrary. Even though the ego vehicle's sensors can detect vehicles farther than 30 metres, the scene focuses on a distance of 15 metres since the driver reaction, while driving in particular situation, is to focalise in shorter ranges [38]. Consequently, vehicles farther than 15 metres are not considered as part of the scene.

Through the definition of the scene, the distinguishing line between the different classes is made clear. The scene, in addition to the ego vehicle, represents the relative positions of each target vehicles detected along the test. As Figure 15 shows, when target vehicles are detected, they are mapped into their discrete space representation. Each of those twelve possible positions included in a scene, represents a valid starting point for a variety of scenarios. Figure 16 gives an example of how a vehicle on the top left position represents the beginning of plenty of different scenarios. The picture shows only two final scenario positions (vehicles in red), however, all positions can be considered as a possible final scenario points. Figure 17 shows how a vehicle can change two lanes, performing a cut-in, with respect to the ego vehicle, ending up on the right lane.

Figure 16 and Figure 17 make clear the connection between the scenario classes and the discrete model.

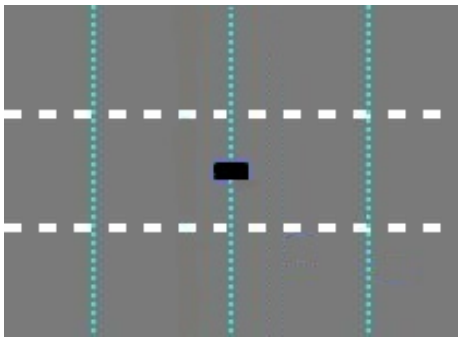


Figure 14: A representation of a scene. The black rectangle represents the ego vehicle driving on the highway: three lanes layout road. The dotted vertical lines show the thresholds defined in Eq. (23).

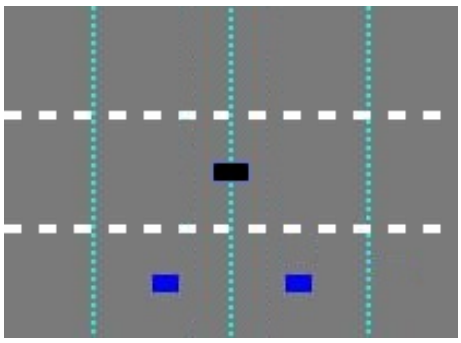


Figure 15: An example of scene with two target vehicles (blue rectangles) detected on the ego vehicle's right (black rectangle).

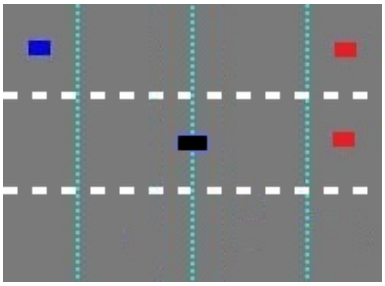


Figure 16: An example of overtaking manoeuvre. The target vehicle (blue rectangle) may conclude the overtaking manoeuvre, as example, in one of those two red rectangles.

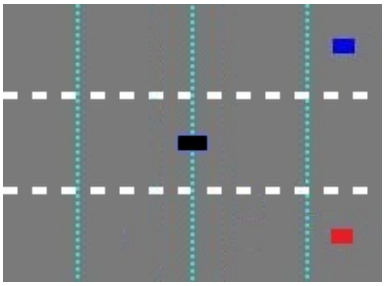


Figure 17: An example of cut-through manoeuvre. The target vehicle may perform a cut-through manoeuvre, ending up where the red rectangle is.

The scenario class has been defined and S of Eq. (5) can be computed by retrieving all observed classes in the database D. An example of how the classes have been identified within the database is given in Section A.

Computing E of Eq. (5), requires the knowledge of how many different scenario classes are possible to encounter under the circumstances delineated in Section 4.1.1. From a practical point of view, according to the discrete model and the expert, six target vehicles are sufficient to completely surround the ego vehicle, therefore $k = 6$. As already mentioned before, all possible starting positions within the scene can be twelve and n represents that quantity.

To calculate E:

$$E = \sum_{i=0}^{k} \binom{n}{i} = \sum_{i=0}^{6} \binom{12}{i} = 2510. \tag{27}$$

Table 1 shows the possible combinations related to each number of vehicle present within the scene, i.e. there are 66 combinations with 2 vehicles located in one of the twelve possible positions.

Table 1: Combinations according to the number of vehicles in the scene. $n = 12$

k	$\binom{n}{k}$
0	1
1	12
2	66
3	220
4	495
5	792
6	924

4.1.3 Weighted function

Exploiting the prior knowledge, this section leads to the definition of the weights that are necessary to compute the completeness so that it reflects as much as possible the assessment's goal. If all 2510 possible combinations have the same importance, then all weights are equal as explain in Section 3.1.3.

Unlike the uniform weights, it might be useful to subgroup the database D and assign them weights according to their importance. The database D is split as follow:

- Group 1: all combinations in which there are 0, 1, 2 and 3 vehicles;
- Group 2: all combinations in which there are 4, 5 and 6 vehicles.

The split aims to associate more importance to the first group because the Property-Damage-Only highlights that injury crashes are more likely to occur in free fluid traffic (i.e. with fewer vehicles around) [39]. Therefore, two weights, respectively w_1 and w_2, are necessary. w_1 refers to the first group, w_2 to the second one and calling back Eq. (2), $L = 2$.

The weights, according to the number of different scenario classes in each group n_1 and n_2, can be computed as:

$$w_1 > w_2 > 0 \tag{28}$$

$$\sum_{j=1}^{L} n_j w_j = \sum_{j=1}^{2} n_j w_j = 1 \tag{29}$$

n_1, number of scenario classes that fall into the first group, is obtained as:

$$n_1 = \sum_{j \in A} \binom{12}{j} = 299, A = \{0, 1, 2, 3\}. \tag{30}$$

n_2 respectively:

$$n_2 = \sum_{j \in B} \binom{12}{j} = 2211, B = \{4, 5, 6\}. \tag{31}$$

Due to the fact that Eqs. (28) and (29) has infinite solutions, many weights can be defined.

Substituting n_j in Eq. (2) it becomes:

$$\sum_{j=1}^{2} n_j w_j = 299 \cdot w_1 + 2211 \cdot w_2 = 1.$$

The overall range of solution for the equation above is:

$$\begin{cases} \frac{1}{2510} < w_1 < \frac{1}{299} \\ w_2 = \frac{1-299 \cdot w_2}{2211} \end{cases} . \tag{32}$$

On one hand, choosing $w_1 = \frac{1}{2510}$ would lead to uniform weights (Eq. (4)), on the other hand, selecting $w_1 = \frac{1}{299}$ would lead to $w_2 = 0$, which is an acceptable solution whether the second group has null importance. It is important to note that by choosing the border solutions, the completeness level may change a lot.

In the following case, the weights such that $w_1 = 2 \cdot w_2$. Therefore, they become:

$$w_1 = 0.000712 \tag{33}$$

$$w_2 = 0.000356 \tag{34}$$

Remark 18. I.e. having observed one scene belonging to group 1 gives a completeness level of 0.000712, whereas a single observed scene belonging to group 2 gives a completeness level of 0.000356. $\Diamond$

4.1.4 SPK's completeness level of D

The database D, taking into account 6 vehicles and according to all delineated boundaries in Section 4.1.1, has detected 23412 items representing 616 scenario classes. As Figure 18 shows, all classes concerning 0, 1 and 2 vehicles are collected while starting from the third column, a light decrease begins, ending up with only few scenario classes including 6 vehicles. The columns in blue speak for the observed scenario classes while the red ones represent all possible combinations computed in Section 4.1.2 and shown in Table 1.

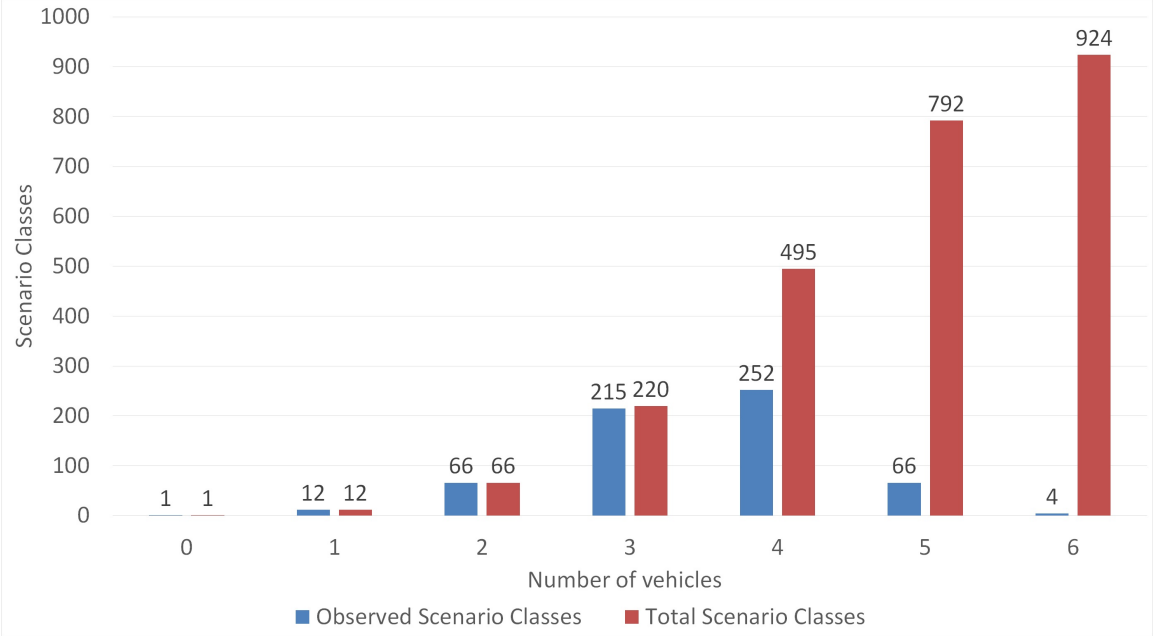


Figure 18: Observed scenario classes.

Calling back Eq. (5), $n_{s,j}$ are the observed scenario classes belonging to the $j-th$ group. They can be computed as follow:

$$n_{s,1} = 1 + 12 + 66 + 215 = 294 \tag{35}$$

$$n_{s,2} = 252 + 66 + 4 = 322 \tag{36}$$

Eventually, the completeness level regarding the database D results as follow:

$$C = \sum_{j=1}^{S} n_{s,j} w_j = 294 \cdot 0.000712 + 322 \cdot 0.000356 = 32.39\% \tag{37}$$

Remark 19. Using the uniform weights,$w = \frac{1}{2510} = 0.0003984$ the completeness level would achieve the following:

$$C = \sum_{j=1}^{S} n_{s,j} w = 294 \cdot 0.0003984 + 322 \cdot 0.0003984 = 24.54\%,$$

which is the exact same result if the completeness level had been calculated as:

$$C = \frac{S}{E} = \frac{294 + 322}{2510} = 24.54\%.$$

$\Diamond$

4.2 Weak Prior Knowledge

The following case study concerns about AD and AVs as the previous one, achieving the scenario classes completeness. The circumstances are slightly different in this case study since estimators will be exploited to compute E and so the completeness level. The observed classes are detected through the discrete model.

The discrete model discloses solely S (how many different scenario classes are present within the database D) since E needs to be estimated.

The Weak Prior Knowledge as Section 3 explained, does not use as much prior knowledge as the Strong Prior Knowledge does. Nonetheless, a discrete model is still necessary to shape the requested completeness level.

No weighted function are considered since E is not known beforehand and uniform weights (Eq. (4)) are applied.

4.2.1 Model discretization

The discrete model connects, as it did for the SPK method Section 4.1.2, the sought completeness with the database D itself by using only the information that matter the most, i.e relative distances and road layout for scenario classes completeness. Once more, it can be seen as the labelling process whose input is the database D and output S, the number of observed scenario classes in D.

The used database D is the same used for the SPK method case study, therefore the gathered information are the same, i.e. velocity, accelerations, yaw rate etc. Yet, the needed ones are the same used in the previous case study as they were the most general features that can be used, especially when low level of prior knowledge is available.

The thresholds are the same ones chosen in Section 4.1.2 but no experts are present to define how many vehicles need to be considered.

Eventually, as Section 3 explained, the discrete model sorts out the scenario classes, returning as outcome S, the number of observed classes within the database D.

4.2.2 Estimations

Chao, Lee and Yang's research [26], [27] developed three estimators Eqs. (12), (13) and (21), pointed out in Section 3.2.2, that are used to compute E.

An easier case is explained first such that the method and reasoning that lay behind the estimators are illustrated. Figure 19 points out 3 as the number of vehicles whose completeness is almost achieved, i.e. 294 scenario classes out of 299. Three vehicles are therefore used to explain the functioning of all estimators.

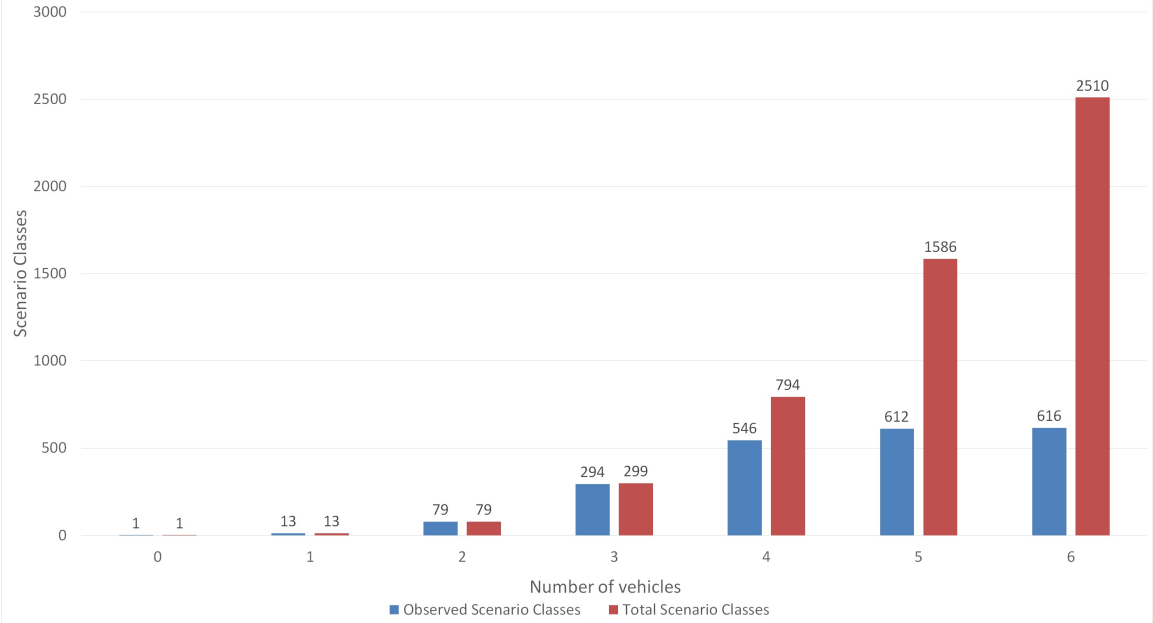


Figure 19: Observed scenario classes versus total number of scenario classes, computed through SPK method.

Figure 20 shows how all estimators N_2, N_3 and $N_{\kappa 10}$ are approaching the same green line marked as *target* which represents the ground truth, i.e. all 299 possible combinations with 3 vehicles (see Section 4.1.2 how to compute the combinations according to the number of vehicles). The chosen value for κ was taken according to [27] and it will be further discussed later when the case study will concern 6 vehicles.

Figures 21 and 22 show the overall behaviour of the estimator's parameters, respectively the occurrences array, f_i, and the *Coefficient of Variations (CV):* $\hat{\gamma}$, $\tilde{\gamma}$, $\tilde{\gamma}_k$. Figure 23 highlights the final part of the estimators performances and their parameters behaviour. In particular, it shows: N_2, N_3 and $N_{\kappa 10}$ achieving the target value (represented by the green line), the occurrences f_i getting close to zero, both $\hat{\gamma}$ (blue line) and $\tilde{\gamma}$ (red line) approaching the assumed correct CV (γ, green line) and γ_k (black line) heading towards zero. These characteristics need to be used as feedback when the combinations number is not known upfront and a judgment on the estimator reliability needs to be drawn.

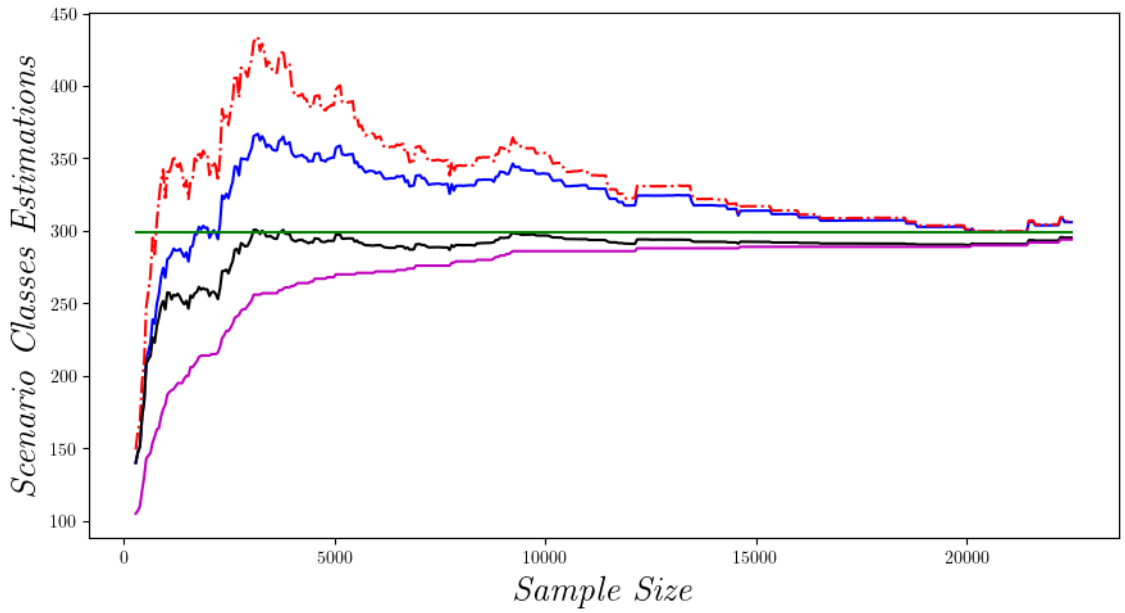


Figure 20: N_2 (blue line), N_3 (red line) and $N_{\kappa 10}$ (black line), estimating the combinations with 3 vehicles detected. Green line speaks for the target.

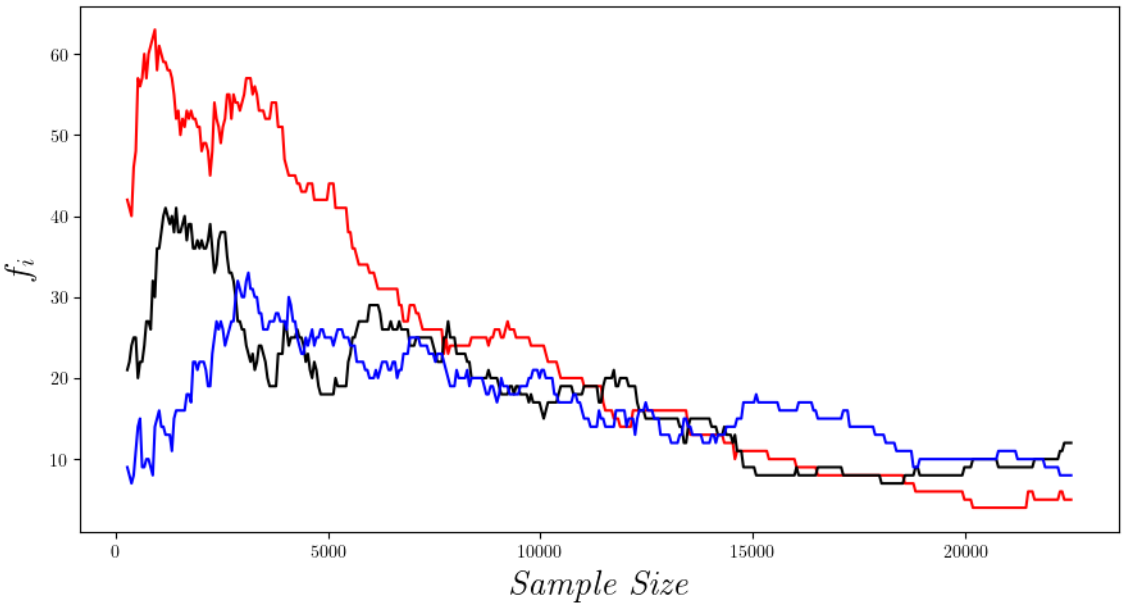
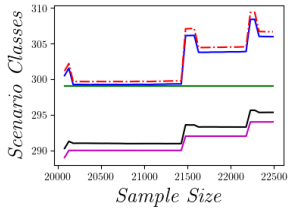
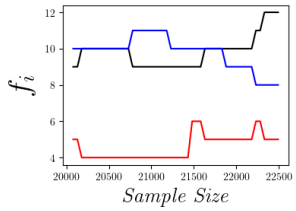


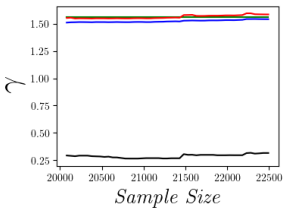
Figure 21: f_1 (red line), f_2 (black line) and f_3 (blue line) occurrences according with the sample size. 3 vehicles detected.



(a) N_2 (blue line), N_3 (red line), N_κ $\kappa = 10$ (black line) and the $target$ (green line).



(b) f_1 (red line), f_2 (black line) and f_3 (blue line).



(c) γ_2 (blue line), γ_3 (red line), γ_κ (black line) and γ (green line).

Figure 23: Zoom-In perspective of the estimations Figure 23a; f_i Figure 23b; related parameters γ Figure 23c.

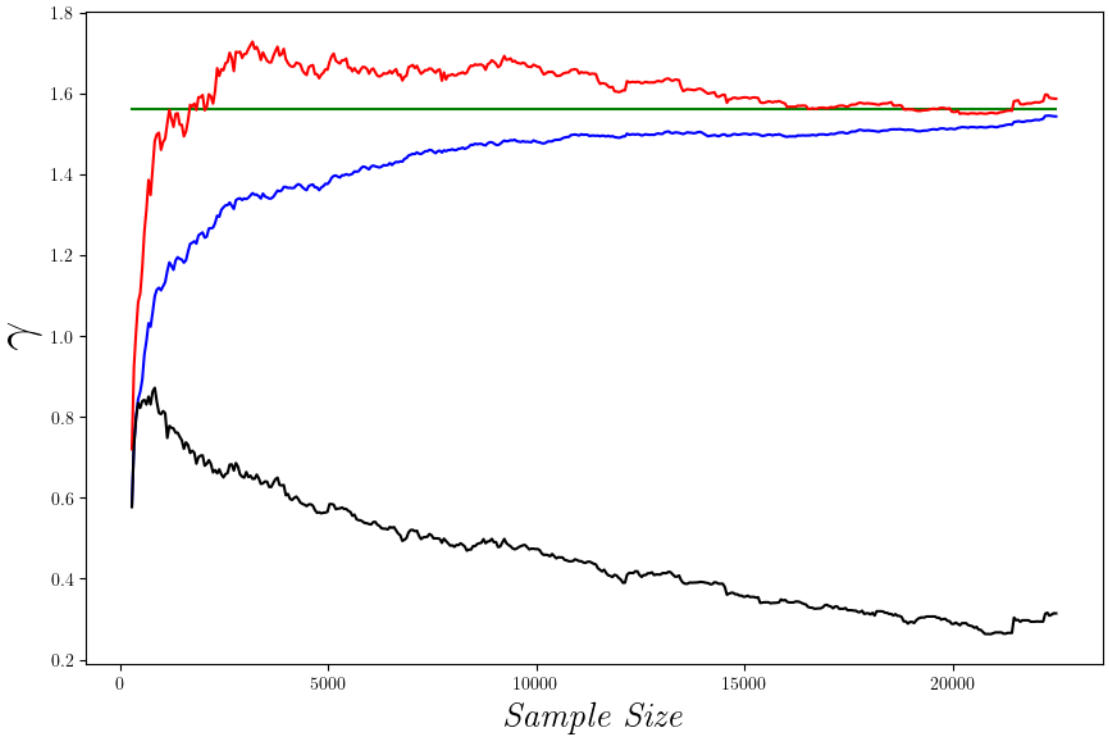


Figure 22: $\hat{\gamma}$ (blue line), $\tilde{\gamma}$ (red line), $\tilde{\gamma}_k$(black line) and γ (green line).

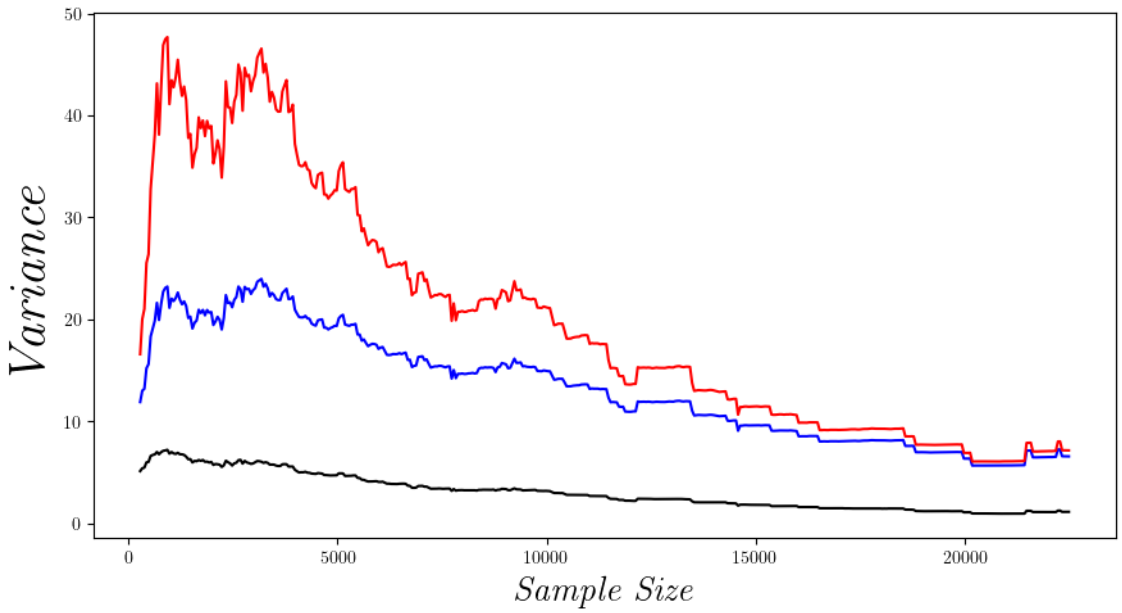


Figure 24: $var(N_2)$ (blue line), $var(N_3)$ (red line) and $var(N_{\kappa 10})$ (black line).

The theory introduced about the estimators in Section 3.2.2 declares that if "infinite" data are gathered the true CV value can be accurately estimated. With a sample size of 23412 and 294 out of 299 observed scenario classes, γ, that is computed according to the following equation:

$$\gamma^2 = \frac{N \sum_{i=1}^{\infty} i(i-1) f_i}{[n(n-1)]} - 1,$$

is assumed at its final value. n is the sample size, $N = 299$ is the number of possible scenario classes.

Remark 20. Note that even though the sum index goes to ∞, f_i is 0 when $i > n$. $\Diamond$

Figure 23c shows that both CVs estimators are heading to its final value, the green line. Furthermore, to prove the proper functioning of the estimators, by analysing the first 3 occurrences array items (f_1, f_2 and f_3), it is important to note that they are all decreasing and approaching zero, as they were supposed to, confirming that the closer the estimators are to the target, the lower are the values of the occurrences f_1, f_2 and f_3.

The last analysed parameter is the variance of the estimators whose values are computed as explained in Eq. (18).

The variance of $N_{\kappa 10}$, as Figure 24 shows, is smaller than the other estimators' variance since the Chao and Yang [27] estimator solely takes into account the first $\kappa = 10$ occurrences. Therefore, a variation in $f_i, i > 10$ does not significantly affect the estimation. On the contrary, N_3, whose variance is the highest amongst all above, has the larger estimated coefficient of variance (γ_{k10}), affecting both outcomes and variances. In addition, as a common property, all variances are decreasing without reaching zero though.

Dealing with 6 vehicles, instead of 3, brings a different problem to be solved:

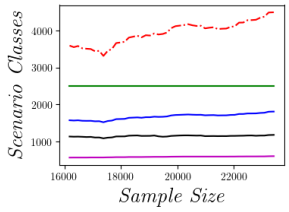
How to understand whether the estimator is getting close to its final estimation?

The given preface with a case study concerning 3 vehicles highlighted all properties that need to be analysed. Thus, the estimator reliability and effectiveness can eventually be drawn when the estimators' convergence is not clear.

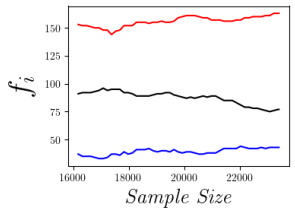
Even though the green line, speaking for the ground truth, can be computed using the SPK method since the discrete model has been chosen equally and the upper bound of the detected vehicles is 6, throughout this case study is assumed as unknown. Therefore an interpretation of the estimators' parameters behaviour is necessary.

As Figure 26 shows, N_3 is extremely high with respect to the other estimations. N_2 and $N_{\kappa 150}$ are close to each other and once more Chao and Yang [27]'s estimator is lower than Chao and Lee [26]'s one. κ has been chosen equal to 150 thus N_κ gets closer to N_2 exploiting the N_2's expected overshooting noticed in Figure 20. A smaller value of κ would have resulted in lower estimations.

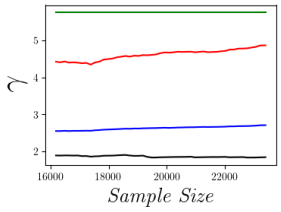
Figures 25, 27 and 28 show how the aforementioned parameters are still increasing. γ_κ and $\hat{\gamma}$ are not converging towards the same value, all f_is are still rising and especially the variance concerning N_3 is extremely large.



(a) N_2 (blue line), N_3 (red line) and N_κ $\kappa = 10$ (black line).



(b) f_1 (red line), f_2 (black line) and f_3 (blue line).



(c) γ_2 (blue line), γ_3 (red line), γ_κ (black line) and γ (green line).

Figure 25: Zoom-In perspective of the estimations Figure 23a; f_i Figure 23b; related parameters γ Figure 23c.

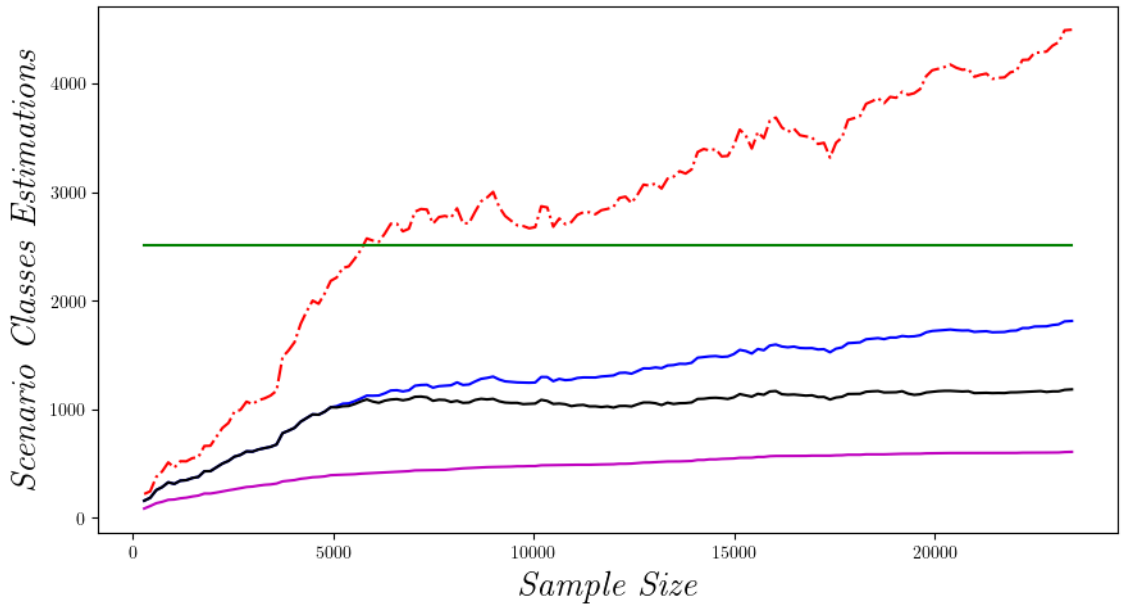


Figure 26: N_2 (blue line),N_3 (red line) and $N_{\kappa 150}$ (black line).

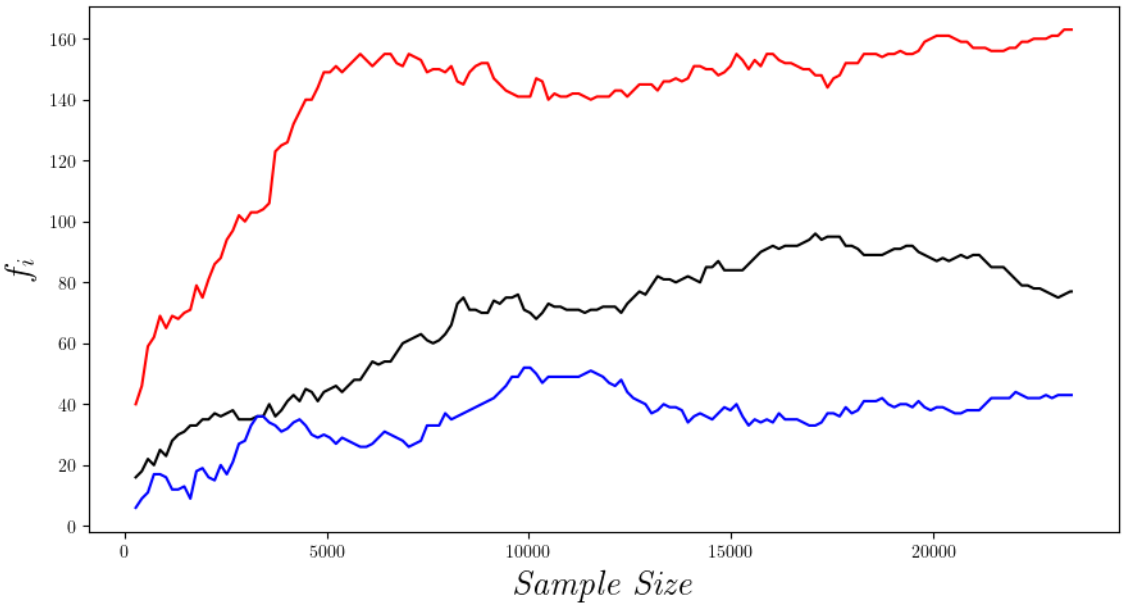


Figure 27: f_1 (red line),f_2 (black line) and f_3 (blue line) occurrences according with the sample size. 6 vehicles detected.

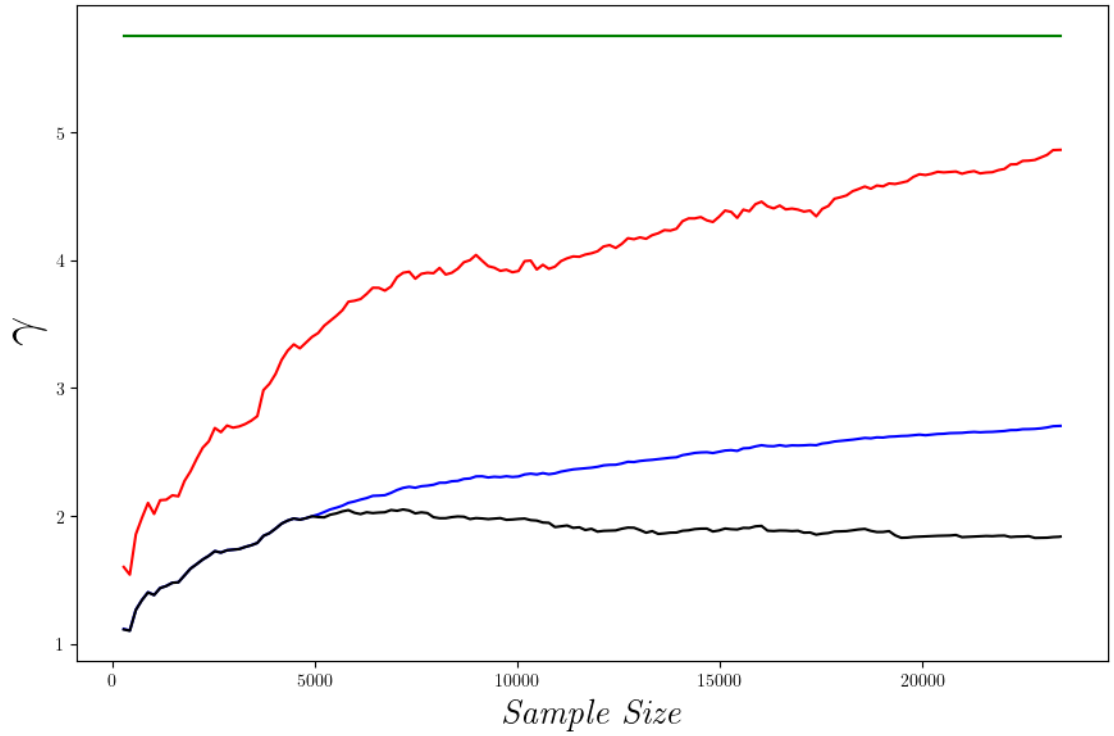


Figure 28: $\hat{\gamma}$ (blue line),$\tilde{\gamma}$ (red line), $\tilde{\gamma}_k$(black line) and γ (green line). 6 vehicles

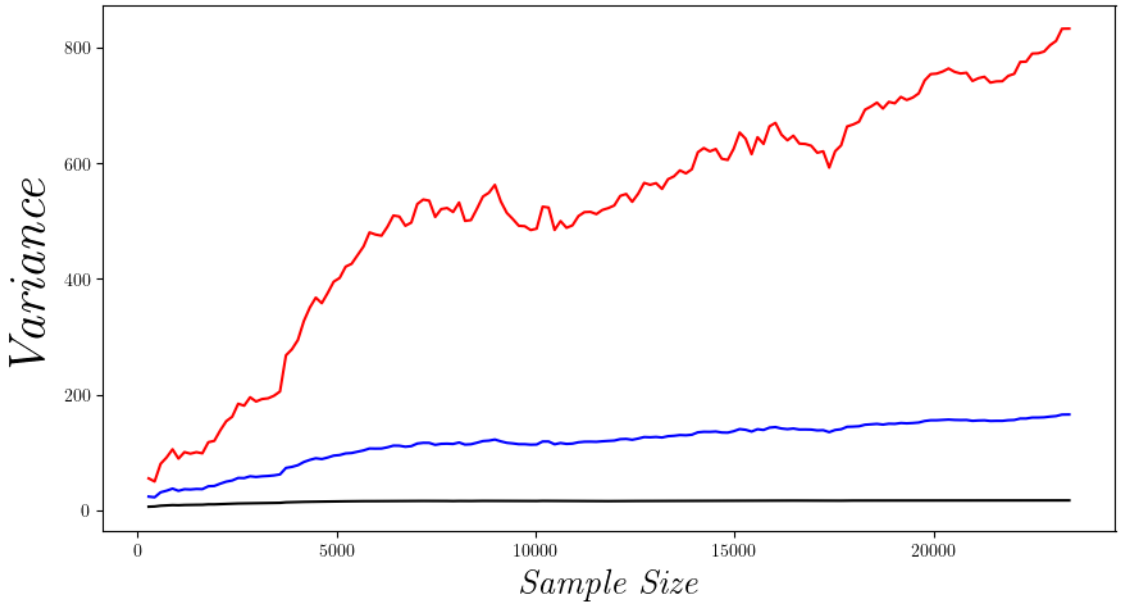


Figure 29: $var(N_2)$ (blue line),$var(N_3)$ (red line) and $var(N_{\kappa 10})$ (black line). 6 vehicles

4.2.3 WPK's completeness level of D

The completeness regarding respectively 3 and 6 vehicles according to all previously studied estimators: N_2, N_3 and N_κ, are shown in Figures 30 and 31. Eq. (5) is used, where S has been computed in Section 4.2.1 and E has been obtained in Section 4.2.2.

The red line in Figure 30 shows a possible positive effect of the overshoot in Figure 20 as having a lower completeness level than the real one (green line).

Black and blue lines in Figure 31 share the completeness level along the first part as the estimations are equal. This behaviour was expected due to the choice of κ.

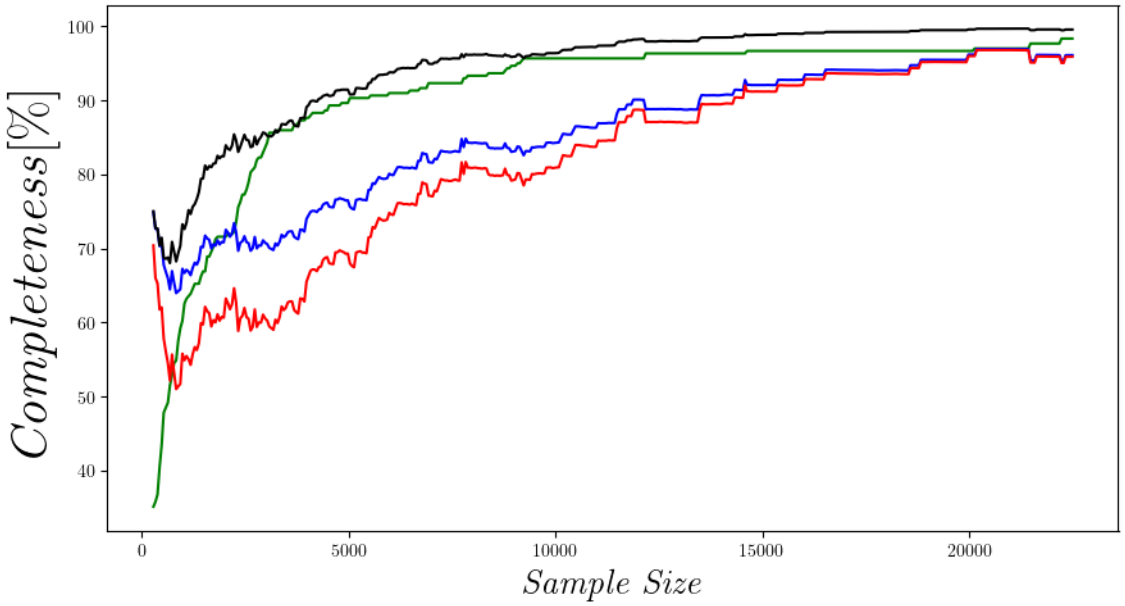


Figure 30: N_2 (blue line),N_3 (red line) and $N_{\kappa 150}$ (black line).

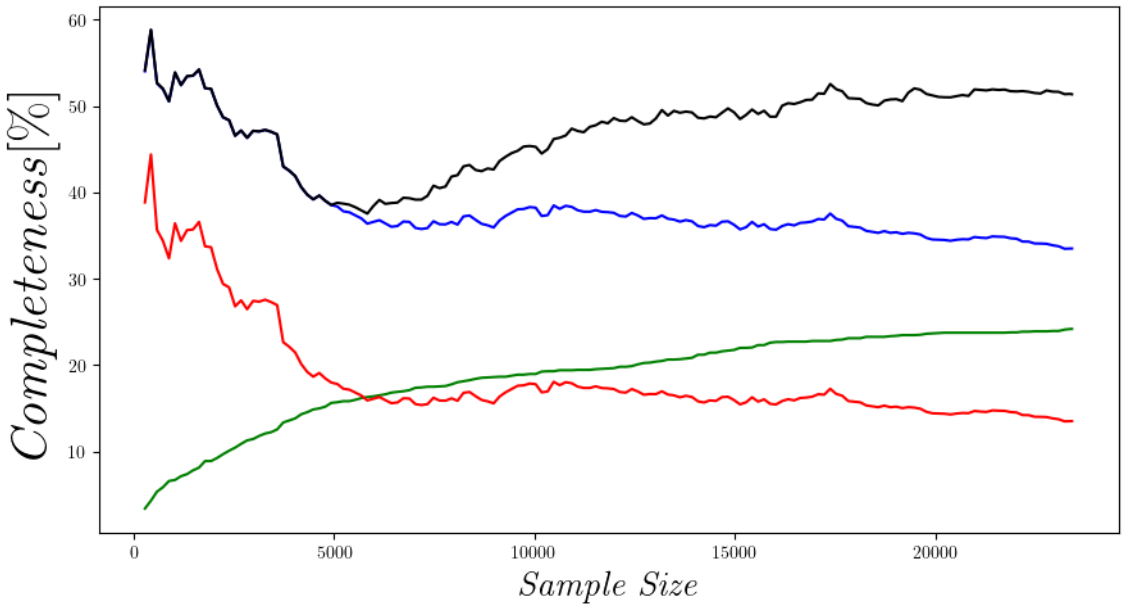


Figure 31: N_2 (blue line),N_3 (red line) and $N_{\kappa 10}$ (black line).

4.3 Discussion

All completeness results are summarised as Table 3 shows whereas all final estimations are shown in Table 2.

Table 2: Estimation results.

Estimator	N. Vehicles	f_1	f_2	f_3	γ	$\tilde{\gamma}$	$\hat{\gamma}$	$\tilde{\gamma}_\kappa$	var	Estimations
N_2	3	5	12	8	1.56	X	1.54	X	6.56	305.71
N_3	3	5	12	8	1.56	1.58	X	X	7.15	306.65
$N_{\kappa 10}$	3	5	12	8	1.56	X	X	0.314	1.13	295.34
N_2	6	163	77	43	5.75	X	2.70	X	165.69	1812.11
N_3	6	163	77	43	5.75	4.86	X	X	832.78	4493.73
$N_{\kappa 150}$	6	163	77	43	5.75	X	X	1.84	17.18	1182.05

Table 3: Completeness levels.

-	N_2	N_3	$N_{\kappa 10}$	N_2	N_3	$N_{\kappa 150}$	SPK_w	$SPK_{\bar{w}}$
Complt. Level [%]	96.08	95.87	99.54	33.5	13.5	51.35	24.54	32.39
N. Vehicles	3	3	3	6	6	6	6	6

$SPK_{\bar{w}}$ represents the outcome from the SPK method using the uniform weights, i.e. 0.0003984. SPK_w speaks as result of using different weights, precisely, $w_1 = 0.000712$ and $w_2 = 0.000356$ where the first group concerns up to 3 vehicles while the second one, which has been considered less important than the first one, takes into account from

4 vehicles to 6, extremes included. The gap between these two results expresses that, according to the chosen weights, the discrete model and all prior knowledge involved, at least a certain number of scenario classes belonging to the first group have been collected. The aim of having different weights, with respect to the uniform ones, is to obtain different completeness level according to the database items, i.e. having collected only scenario classes that belong to the second group would have given a lower completeness than the uniform one. On the other hand, having a higher level of completeness with respect to the uniform one, can draw as conclusion that at least, some scenario classes belonging to the first group are represented within the database.

Details regarding the number of classes detected in each group can be drawn by computing the different percentage range between $SPK_{\bar{w}}$ and SPK_w. The percentage difference, defined as $D\% = SPK_w - SPK_{\bar{w}}$, has a span of: 18.74%, raging from -9.37% to $+9.37\%$, can be computed as:

$$\begin{aligned} \max &= (n_1 \cdot w_1 - n_1 \cdot w) \cdot 100 = 9.37\% \\ \min &= (n_2 \cdot w_2 - n_2 \cdot w) \cdot 100 = -9.37\%. \end{aligned}$$

Having a positive $D\%$ means that the proportion between the observed scenario classes that belong to the most important subgroup is higher than the observed scenario classes of the least important subgroup, i.e. $\frac{n_{s,1}}{n_1} > \frac{n_{s,2}}{n_2}$.

The percentage of completeness graphically shown in Figure 31 and Table 3, pretending to not being aware of the ground truth, leads to keep collecting more representative data since the highest percentage is barely over 50%. Further analysing the results, by looking at the previously studied parameters in Figure 25, the computed percentage of completeness suggests that the data is far from being complete since all parameters are still either growing or far away from the expected behaviour earlier studied with 3 vehicles.

Figure 27 shows how all occurrences are still increasing; Figure 28 has its red and blue line approaching each other but still far apart. Last symptom saying that the collected data is not enough, is shown in Figure 29. The plot shows a vast variance related to N_3, which could be foreseen from the case study with 3 vehicles, but yet, even though the variance of N_2 is by far lower than the N_3's one, it is still almost 166 and, more important, the slope is positive, meaning that the N_2's variance is going to increase its value even more.

It is important to note how the largest estimation, N_3, ends up being the most conservative completeness level amongst the other ones. Under certain circumstances a more conservative level of completeness might be preferred over an overestimating one.

Eventually, comparing the case where solely 3 vehicles were involved and the one with 6

vehicles, the latter has collected almost 25% of the total number of scenario classes while the first has more than 98%. Even though N_2 seems to be close to the real completeness, an overshoot of its estimation might occur like happened in Figure 20. On the contrary, whether the behaviour of N_κ is going to be the same as Figure 20 shows, it will not overshoot except with larger value of κ. In the latter case N_κ would follow N_2 ending up overestimating E and therefore, underestimating the completeness level.

5 Conclusion

Recently, mobility and transportation have improved a lot, allowing to connect all the world faster and cheaper. The ongoing research in traffic and transport context are facing the fourth level of autonomous driving. Nonetheless, in parallel to these breakthroughs, the vehicle safety has rose up as problem to overcome, especially when the human-car interactions are slowly getting outside of the driving loop control.

Scenario-based approach paves the path towards the needed vehicle safety claims, however, resources, time and costs need to be considered for gathering a representative database. On the one hand, insufficient data may bring to misleading results while on the other hand excessive data lead to waste resources, therefore, questions about how to measure the completeness of the database began to rise interests in various fields.

This thesis proposed two methods named Strong Prior Knwoledge (SPK) and Weak Prior Knwoledge (WPK). Solely three general assumptions are assumed with respect to the scenario classes: they do not change over time, they are noise-free and they are identically and randomly observed.

The SPK method exploited a broaden overview of the problem and starting from the data collection process through the discrete model and the weighted function made use of the available prior knowledge. The method explained how the completeness problem meets the collection process' goals through the prior knowledge. By using the prior knowledge the method was able to compute how many different scenario classes can be encountered by an AV once that the discrete model was defined and therefore, obtain the completeness level.

As second method, a new statistical approach called Weak Prior Knowledge, assuming that either not enough prior knowledge was available or the discrete models were too complex to achieve the number of combinations upfront. The number of possible scenario classes was estimated by using estimators exploited in biology to evaluate the number of different living animal species. Nonparametric estimators were used to estimate the number of possible scenario classes without using prior knowledge of scenario class distributions but solely considering the occurrences of each observed scenario class.

A real-world traffic scenarios database had been used as case study to demonstrate the functioning of both methods, WPK and SPK. Eventually, both methods computed their level of completeness regarding the database concluding that not all possible scenario classes were observed.

Both proposed methods were built up to help future decisions on data collection with close attention to the prior knowledge and the discrete model used. Different discrete models and/or prior knowledge, however, may lead to diverse completeness levels.

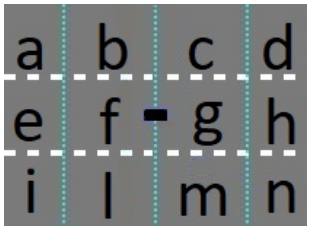
This thesis answered the question of measuring the degree of completeness regarding

the scenario classes. However, the extensions of this work to also address the question of the degree of completeness of all scenarios that belong to a particular scenario class, remains future work.

Appendices

A Coding

The scene explained throughout the thesis are represented and handled differently in the code. The code, which has been built in Python environment, foresees a transformation from raw data (real-world traffic data) to vector strings. In particular, the code creates a vector whose positions represent either the presence or the absence of any vehicle. The vector is composed of binary information signalising the presence/absence of the vehicle. Figure 32 shows both representations.



(a) Graphical overview of a scene. Ego vehicle in the centre and each letter corresponds to a possible starting/ending position.

a	b	c	d	e	f	g	h	i	l	m	n

(b) How the scene on the left is represented within the code. Each letter can be either 0 *or* 1.

Figure 32: How the scene are represented in the code.

How the discrete model is built affects the number of detected classes. The code associates the detected vehicles to the most affine class when a proper discretization does not exist. As example, if the discrete model involves only a three lanes road layout while the real road is a four-road lanes, the vehicles detected in the most left lane will be included in the next right lane. This may represent a problem because diverse scenario classes would be considered as similar ones. Furthermore, changing the number of vehicles to detect, n, may affect the number of observed classes. The labelling process that associates an ID to each vehicle stops when n is achieved. It may occur that if the labelling process did not stop and consequently more vehicles were detected, more classes would be identified. This is caused by the reduction of the region of interest expressed as $-15 \leq d \leq 15$ respect to the sensor detection range.

Computationally speaking, computing the variance stated in Eq. (18), it is important to note that considering $f_i = 0$ or not consider it at all, does not affect the variance. According to that, the variance computation uses only the first 1500 items of the f array since, analysing the data and with respect to the database D, no classes have been detected more than 1500 times.

References

[1] K. Bengler, K. Dietmayer, B. Färber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives", *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 4, pp. 6–22, 2014.

[2] D. U.S. Department of Transportation Washington, "Bureau of transportation statistics, 2014a. motorcycle rider (operator) safety data", 2016.

[3] K. Bimbraw, "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology", in *12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, vol. 1, 2015, pp. 191–198.

[4] "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles", SAE International, Tech. Rep. J3016, 2018.

[5] E. de Gelder, J.-P. Paardekooper, O. Op den Camp, and B. De Schutter, "Safety assessment of automated vehicles: How to determine whether we have collected enough field data?", *Traffic injury prevention*, vol. 20, no. sup1, S162–S170, 2019.

[6] J. Ploeg, E. d. Gelder, M. Slavik, E. Querner, T. Webster, and N. d. Boer, "Scenario-based safety assessment framework for automated vehicles", in *Proceedings of the 16th ITS Asia-Pacific Forum, 713-726*, 2018.

[7] H. Elrofai, J.-P. Paardekooper, E. de Gelder, S. Kalisvaart, and O. O. den Camp, "Scenario-based safety validation of connected and automated driving", 2018.

[8] E. de Gelder and J.-P. Paardekooper, "Assessment of automated driving systems using real-life scenarios", in *2017 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2017, pp. 589–594.

[9] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data", *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016.

[10] S. Alvarez, Y. Page, U. Sander, F. Fahrenkrog, T. Helmer, O. Jung, T. Hermitte, M. Düering, S. Döering, and O. Op den Camp, "Prospective effectiveness assessment of adas and active safety systems via virtual simulation: A review of the current practices", in *25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, 2017.

[11] S. Geyer, M. Baltzer, B. Franz, S. Hakuli, M. Kauer, M. Kienle, S. Meier, T. Weißgerber, K. Bengler, R. Bruder, *et al.*, "Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance", *IET Intelligent Transport Systems*, vol. 8, no. 3, pp. 183–189, 2013.

[12] C. Liu, A. Talaei-Khoei, D. Zowghi, and J. Daniel, "Data completeness in healthcare: A literature survey", *Pacific Asia Journal of the Association for Information Systems*, vol. 9, no. 2, 2017.

[13] A. L. Lamberg, D. Cronin-Fenton, and A. B. Olesen, "Registration in the danish regional nonmelanoma skin cancer dermatology database: Completeness of registration and accuracy of key variables", *Clinical epidemiology*, vol. 2, p. 123, 2010.

[14] E. S. Berner, R. K. Kasiraman, F. Yu, M. N. Ray, and T. K. Houston, "Data quality in the outpatient setting: Impact on clinical decision support systems", in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2005, 2005, p. 41.

[15] N. A. Emran, S. Embury, P. Missier, M. N. M. Isa, and A. K. Muda, "Measuring data completeness for microbial genomics database", in *Asian Conference on Intelligent Information and Database Systems*, Springer, 2013, pp. 186–195.

[16] F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet, "Genotyping errors: Causes, consequences and solutions", *Nature Reviews Genetics*, vol. 6, no. 11, p. 847, 2005.

[17] A. Gandolfi and S. Chelluri, "Nonparametric estimations about species not observed in a random sample", *Milan Journal of Mathematics*, vol. 72, no. 1, pp. 81–105, 2004.

[18] R. N. Lockwood, J. C. Schneider, R. N. Lockwood, and J. C. Schneider, *Chapter 7: Stream fish population estimates by mark-and-recapture and depletion methods*, 2000.

[19] J. Bunge and M. Fitzpatrick, "Estimating the number of species: A review", *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 364–373, 1993.

[20] A. Chao, "Nonparametric estimation of the number of classes in a population", *Scandinavian Journal of statistics*, pp. 265–270, 1984.

[21] K. P. Burnham and W. S. Overton, "Robust estimation of population size when capture probabilities vary among animals", *Ecology*, vol. 60, no. 5, pp. 927–936, 1979.

[22] B. Harris, "Statistical inference in the classical occupancy problem unbiased estimation of the number of classes", *Journal of the American Statistical Association*, vol. 63, no. 323, pp. 837–847, 1968.

[23] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?", *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[24] W. Wang, C. Liu, and D. Zhao, "How much data are enough? a statistical approach with case study on longitudinal driving behavior", *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 2, pp. 85–98, 2017.

[25] F. Hauer, T. Schmidt, B. Holzmüller, and A. Pretschner, "Did we test all scenarios for automated and autonomous driving systems?", 2019.

[26] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage", *Journal of the American statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[27] A. Chao and M. C. Yang, "Stopping rules and estimation for recapture debugging with unequal failure rates", *Biometrika*, vol. 80, no. 1, pp. 193–201, 1993.

[28] P. E. Project, "Proactive safety for pedestrian and cyclists", 2018.

[29] H. Yang, B. Van Dongen, A. Ter Hofstede, M. Wynn, and J. Wang, "Estimating completeness of event logs", *BPM Center Report*, vol. 12, 2012.

[30] M. Egger and G. D. Smith, *Misleading meta-analysis*, 1995.

[31] M. E. Lladser, "Prediction of unseen proportions in urn models with restricted sampling", in *2009 Proceedings of the Sixth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, SIAM, 2009, pp. 85–91.

[32] K. P. Burnham and W. S. Overton, "Estimation of the size of a closed population when capture probabilities vary among animals", *Biometrika*, vol. 65, no. 3, pp. 625–633, 1978.

[33] P. Flajolet, D. Gardy, and L. Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search", *Discrete Applied Mathematics*, vol. 39, no. 3, pp. 207–229, 1992.

[34] J. Bunge and K. Barger, "Parametric models for estimating the number of classes", *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 50, no. 6, pp. 971–982, 2008.

[35] F. Olken and D. Rotem, "Simple random sampling from relational databases", 1986.

[36] J. N. Darroch, "The multiple-recapture census: I. estimation of a closed population", *Biometrika*, vol. 45, no. 3/4, pp. 343–359, 1958.

[37] J.-P. Paardekooper, S. Montfort, J. Manders, J. Goos, E. de Gelder, O. Op den Camp, A. Bracquemond, and G. Thiolon, "Automatic identification of critical scenarios in a public dataset of 6000 km of public-road driving", in *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2019.

[38] E. Lehtonen, O. Lappi, H. Kotkanen, and H. Summala, "Look-ahead fixations in curve driving", *Ergonomics*, vol. 56, no. 1, pp. 34–44, 2013.

[39] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states", *Traffic injury prevention*, vol. 16, no. 1, pp. 28–35, 2015.