AGH University of Science and Technology

Faculty of Electrical Engineering, Automatics, Computer Science
and Biomedical Engineering
DEPARTMENT OF APPLIED COMPUTER SCIENCE

Sorbonne Université

École doctorale Informatique, Télécommunications et Électronique
LABORATOIRE D'INFORMATIQUE DE PARIS 6

PHD THESIS

MARCIN LENART

SENSOR INFORMATION SCORING FOR DECISION-AID SYSTEMS IN RAILWAY DOMAIN

SUPERVISOR:
dr hab. Andrzej Bielecki, prof. AGH
CO-SUPERVISOR:
dr hab. Marie-Jeanne Lesot, assoc. prof. LIP6, Sorbonne Université
ADVISORS:
dr Teodora Petrisor, research engineer Thales
dr Adrien Revault d'Allonnes, assoc. prof. LIASD, University Paris VIII

Cracow 2019

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
KATEDRA INFORMATYKI STOSOWANEJ

Sorbonne Université
École doctorale Informatique, Télécommunications et Électronique
LABORATOIRE D'INFORMATIQUE DE PARIS 6

ROZPRAWA DOKTORSKA

MARCIN LENART

OCENA JAKOŚCI INFORMACJI DLA SYSTEMÓW DECYZYJNYCH W DOMENIE KOLEJOWEJ

PROMOTOR:
dr hab. Andrzej Bielecki, prof. AGH
PROMOTOR POMOCNICZY:
dr hab. Marie-Jeanne Lesot, assoc. prof. LIP6, Sorbonne Université
DORADCY:
dr Teodora Petrisor, research engineer Thales
dr Adrien Revault d'Allonnes, assoc. prof. LIASD, University Paris VIII

Kraków 2019

Acknowledgements

This PhD thesis has been financed by Thales Polska. I am very grateful to them for allowing this research to happen. I would like to acknowledge a Thales expert: Piotr Piotrowski who supported my experiments with deep knowledge and advise.

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would like to thank all my advisor: Marie-Jeanne Lesot, Andrzej Bielecki, Teodora Petrisor and Adrien Revault d'Allonnes, whose expertise was invaluable in the formulating of the research topic and methodology in particular.

I would like to acknowledge my colleagues from both universities AGH and Sorbonne as well as from Thales company for their wonderful collaboration. You supported me greatly and were always willing to help me.

In addition, I would like to thank my family for their wise counsel and sympathetic ear. You are always there for me. Finally, there are my friends, who were of great support in deliberating over our problems and findings, as well as providing a happy distraction to rest my mind outside of my research.

Summary

In this thesis, we investigate the problem of assessing information quality produced by sensors. Information quality is an abstract concept which is considered important in many fields as high-quality information is necessary for decision-making systems. Its assessment highly depends on the type of information, its context and considered domain. It is usually decomposed into different criteria, called dimensions, that allow to capture and combine different aspects of a piece of information. This thesis focuses on the case of information produced by sensors, i.e. devices measuring an aspect of reality and translating it into a digital value. Indeed, sensors, usually used in networks, do not always provide correct information and the scoring of this information is needed. This thesis proposes to exploit the sensor specificity to define a dedicated, and yet generic, scoring method.

Existing approaches for information scoring in the case of sensors are usually based on ground truth or meta-information, which significantly limits their genericity: they are often difficult to obtain and make the approaches appropriate only for specific sensors, exploiting their unique characteristics.

We propose an approach that deals with these difficulties by providing a model designed to be sensor-generic, not dependent on ground truth and dependent only on easy-to-access meta-information, exploiting only attributes shared among the majority of sensors. The proposed model is called ReCLiC from the four dimensions that it considers: Reliability, Competence, Likelihood and Credibility.

Informally, the ReCLiC model takes as input a log file of sensor entries and aims at attaching each log entry with a numerical evaluation of quality of this entry: this quality is understood as the trust that can be put in the message content of the log entry, measured by considering the source, the content and the context of this message, which are the three main components defining a piece of information. We discuss in depth the requirements of the four proposed dimensions on which ReCLiC relies and propose motivated definitions for each of them. Furthermore, we propose an implementation of the generic ReCLiC definition to a real case, for a specific sensor in the railway signalling domain: we discuss the form of the four dimensions for this case and perform a formal study of the information scoring behaviour, analysing each dimension separately.

The proposed implementation of the ReCLiC model is experimentally validated using realistic simulated data created from a real dataset in the railway domain. The proposed experimental protocol allows to control various quality issues as well as their quantity, in four distinct scenarios of problematic log files. This experimental study that includes a study of the parameters shows that the proposed ReCLiC model has the desired behaviour and in particular the ability to assign low trust scores to the simulated noisy entries.

Finally, the ReCLiC model is used to analyse a real dataset where quality problems are detected and discussed. A new visualisation method is proposed to show multiple trust scores from many sensors at the same time. This visualisation allows to observe trust propagation which shows how low-quality messages can impact other information. In addition, the notion of trust dynamics is introduced and analysed based on this example.

Keywords: Information quality, data quality, quality scoring, trust, sensors, reliability, competence, likelihood, credibility, railway.

Résumé

Dans cette thèse, nous examinons le problème de l'évaluation de la qualité d'information produite par des capteurs. La qualité de l'information est un concept d'une grande importance dans de nombreux domaines, car la prise en compte d'information de grande qualité est nécessaire pour les systèmes d'aide à la décision. Leur évaluation, ou cotation, dépend du type des informations, de leur contexte ainsi que du domaine considéré. La qualité de l'information est le plus souvent décomposée en plusieurs critères, appelés dimensions, qui permettent de capturer et de combiner différents aspects d'une information. Cette thèse considère le cas d'informations produites par des capteurs, c'est-à-dire des systèmes qui mesurent un aspect de la réalité et le transforment en une valeur numérique. En effet, les capteurs, le plus souvent utilisés en réseaux, ne fournissent pas toujours une information correcte et l'évaluation de sa qualité est nécessaire. Cette thèse propose d'exploiter les spécificités des capteurs pour définir un modèle de cotation d'information dédié, et cependant générique.

Les approches existantes pour la cotation d'information produite par les capteurs reposent souvent sur une vérité-terrain ou des méta-informations, ce qui restreint significativement leur généricité : vérité-terrain et méta-informations sont souvent difficiles à obtenir et rendent les approches appropriées uniquement pour des capteurs spécifiques, en exploitant des caractéristiques propres.

Nous proposons une approche qui traite ces difficultés en définissant un modèle qui ne fait pas d'hypothèse sur le capteur considéré, ne requiert pas de vérité-terrain et dépend seulement de méta-informations aisées à obtenir, qui exploitent uniquement des attributs partagés par la plupart des capteurs. Le modèle que nous proposons est appelé ReCLiC, du fait des noms en anglais des quatre dimensions sur lesquelles il repose : fiabilité, compétence, vraisemblance et crédibilité, soit Reliability, Competence, Likelihood et Credibility.

De façon informelle, le modèle ReCLiC prend en entrée les messages fournis par un ensemble de capteurs et vise à enrichir chacun des messages d'une évaluation numérique de sa qualité. Cette qualité est comprise comme la confiance qui peut être mise dans ce message. L'évaluation de la qualité dépend de la source, le contenu et le contexte du message, qui sont les trois composantes principales d'une information. Nous discutons en détails les contraintes et propriétés souhaitées des quatre dimensions sur lesquelles ReCLiC repose et nous proposons des définitions motivées pour chacune d'entre elles. De plus, nous proposons une implémentation de la définition générique de ReCLiC pour un problème réel, pour un capteur spécifique dans le domaine de la signalisation ferroviaire : nous discutons de leur forme pour cette application et nous effectuons une analyse théorique du comportement du modèle de cotation d'information auquel elles conduisent, en examinant chaque dimension séparément.

L'implémentation proposée de ReCLiC est validée expérimentalement en utilisant des données simulées réalistes, créées à partir d'une base de données ferroviaires réelles, fournie par

le partenaire industriel de la thèse, Thales. Le protocole expérimental que nous proposons permet de contrôler les problèmes de qualité introduits ainsi que leur nombre, selon quatre scénarios distincts. Cette étude expérimentale porte également sur les paramètres du modèle proposé ReCLiC et permet de montrer qu'il offre le comportement souhaité : en particulier, il possède la capacité d'affecter des scores de confiance faibles aux messages bruités simulés.

Enfin le modèle ReCLiC est mis en œuvre pour analyser la base de données ferroviaires réelles. Une nouvelle méthode de visualisation est proposée, pour représenter graphiquement de multiples scores de confiance associés aux messages de multiples capteurs simultanément. Cette visualisation permet d'observer un phénomène de propagation de confiance qui montre comment des messages de faible qualité influencent d'autres messages. De plus, la notion de dynamique de la confiance est introduite et analysée sur ces données.

Mots-clés : qualité de l'information, qualité des données, cotation d'information, confiance, capteurs, fiabilité, compétence, vraisemblance, crédibilité, donnés ferroviaires.

Streszczenie

W niniejszej dysertacji badany jest problem oceny jakości informacji produkowanej przez urządzenia pomiarowe. Jakość informacji jest pojęciem abstrakcyjnym, które uważa się za istotne w wielu dziedzinach jako że informacje dobrej jakości są niezbędne do działania wielu różnych systemów decyzyjnych. Ocena jakości informacji w dużej mierze zależy od rodzaju informacji, jej kontekstu i rozważanej dziedziny. Zwykle ocena jakości informacji polega na jej rozłożeniu na pojedyncze elementy, zwane wymiarami, które pozwalają uchwycić i połączyć różne aspekty informacji. Niniejsza praca koncentruje się na przypadku informacji produkowanych przez urządzenia pomiarowe, tj. urządzenia, które mierzą fragment rzeczywistości i przekształcają ją na wartość cyfrową. Istotnie, urządzenia pomiarowe, zwykle stosowane w większych grupach, nie zawsze produkują poprawne informacje i konieczna jest ich ocena. W niniejszej dysertacji proponujemy wykorzystanie specyficznych aspektów urządzeń pomiarowych w celu zdefiniowania dedykowanej, ale także ogólnej metody oceniania.

Istniejące propozycje oceny jakości informacji produkowanych przez urządzenia pomiarowe są w większości oparte na wykorzystywaniu danych uczących lub meta-informacji, co znacznie ogranicza ich ogólność: informacje te są często trudne do uzyskania i sprawiają, że metody te mogą być wykorzystywane tylko w jednej sytuacji, dla specyficznie określonych urządzeń pomiarowych, wykorzystując ich unikalne cechy.

Zaproponowany w pracy model działa niezależnie od obecności danych uczących oraz specyficznych meta-informacji o danym urządzeniu. Model ten jest zaprojektowany tak, aby mógł zostać wykorzystany w przypadku informacji produkowanych przez każdy rodzaj urządzenia pomiarowego wykorzystując do działania jedynie atrybuty wspólne dla większości urządzeń i łatwo dostępne meta-informacje. Proponowany model nosi nazwę ReCLiC od angielskich nazw czterech wykorzystywanych wymiarów: niezawodność, kompetencja, prawdopodobieństwo i wiarygodność.

Model ReCLiC przyjmuje na wejście wiadomości z bazy danych i ma na celu przyłączenie do każdej wiadomości numeryczną wartość która jest oceną jakości tej wiadomości: jakość ta jest rozumiana jako zaufanie, które można pokładać w treści tej wiadomości. Zaufanie to jest mierzone poprzez ocenę źródła, treści i kontekstu wiadomości, które są trzema głównymi składnikami definiującymi informację. W pracy dokładnie omawiamy wymagania czterech proponowanych wymiarów, na których opiera się ReCLiC i proponujemy definicje dla każdego z nich. Ponadto, w pracy proponowana jest implementacja ogólnej definicji ReCLiC do rzeczywistego przypadku dla konkretnego urządzenia kolejowego: omawiamy definicje czterech wymiarów dla tego urządzenia i przeprowadzamy formalne badanie proponowanych definicji, analizując potencjalne zmiany poziomu zaufania na podstawie zmian każdego z czterech wymiarów.

Proponowana implementacja ReCLiC jest weryfikowana eksperymentalnie przy użyciu

danych symulowanych opartych na rzeczywistej bazie danych z dziedziny kolejowej. Zaproponowany eksperyment pozwala na kontrolę różnych potencjalnych problemów związanych z jakością informacji oraz ich ilości przy użyciu czterech różnych scenariuszy. Ta eksperymentalna walidacja, która dodatkowo obejmuje modyfikację różnych parametrów, pokazuje, że proponowany model ReCLiC właściwie ocenia jakość informacji, a w szczególności zdolność do przypisywania niskich poziomów zaufania dla symulowanych wiadomości.

Dodatkowo, model ReCLiC jest wykorzystany do analizy prawdziwego zbioru danych, w którym wykrywane i omawiane są problemy z jakością informacji. W tym celu została zaproponowana nowa metoda wizualizacji która ma za zadanie zobrazować wiele poziomów zaufania dla wielu urządzeń jednocześnie. Pozwala ona także zaobserwować propagację niskiego poziomu zaufania, która pokazuje w jaki sposób wiadomości niskiej jakości mogą wpływać na jakość innych informacji. Dodatkowo wprowadzono i omówiono pojęcie dynamiki zaufania na podstawie tego przykładu.

Słowa kluczowe: Jakość informacji, jakość danych, ocena jakości, zaufanie, urządzenia pomiarowe, niezawodność, kompetencja, prawdopodobieństwo, niezawodność, kolej.

Contents

1. Introduction ... 17
1.1. Motivation and goals of the thesis ... 17
1.2. Structure of the thesis ... 18
2. Foundations and key concepts ... 21
2.1. Information scoring: measuring information quality ... 21
2.1.1. Data versus information ... 21
2.1.2. Reminder about Data Quality ... 23
2.1.3. Information quality ... 25
2.1.4. Focus on a trust model ... 27
2.2. Information quality for sensor measurements ... 28
2.2.1. Sensor definition and characteristics ... 29
2.2.2. Existing models ... 30
2.3. Dynamic analysis: quality evolution ... 33
2.4. Summary ... 34
3. Dynamic trust scoring for sensors: the proposed ReCLiC model ... 35
3.1. Goals and requirements of the proposed model ... 35
3.1.1. Desired characteristics ... 35
3.1.2. Considered data input and notation ... 36
3.1.3. Meta-information input ... 37
3.2. Overview of the proposed approach ... 39
3.2.1. General principle ... 39
3.2.2. Graphical representation of the ReCLiC model ... 41
3.3. Reliability: initial source evaluation ... 42
3.3.1. Discussion ... 42
3.3.2. Proposed definition ... 43
3.4. Competence: a refined source evaluation ... 44
3.4.1. Discussion ... 45

3.4.2. Proposed definition 46
3.5. Likelihood: temporal confirmation 47
3.5.1. Discussion 47
3.5.2. Proposed definition 48
3.6. Credibility: spatial confirmation 49
3.6.1. General definition 50
3.6.2. Sets of candidate sources 50
3.6.3. Sets of confirming and invalidating messages 51
3.6.4. Aggregation 52
3.6.5. The final credibility aggregation 55
3.7. Trust 57
3.7.1. Discussion 57
3.7.2. The final trust aggregation 59
3.8. Summary 60
4. ReCLiC implementation for the railway signalling domain: the case of Axle Counters 61
4.1. Devices in the railway signalling domain 61
4.1.1. Overview 61
4.1.2. Axle Counters 63
4.1.3. Motivation for scoring trust in Axle Counters 65
4.2. Adaptation of the ReCLiC model to the AC case 66
4.2.1. Reliability 66
4.2.2. Competence 67
4.2.3. Likelihood 68
4.2.4. Credibility 71
4.2.5. Trust 75
4.3. Formal study of trust evolution and its components 76
4.3.1. Reliability analysis 77
4.3.2. Competence analysis 78
4.3.3. Likelihood analysis 78
4.3.4. Credibility analysis 79
4.4. Summary 80
5. Experimental study on realistic simulated data 81
5.1. Motivation 81
5.2. Proposed scenarios for the generation of realistic simulated data 82

5.2.1. Overview 82
5.2.2. Uniform noise 83
5.2.3. Burst noise 83
5.2.4. Random message injection 84
5.2.5. Non-existent message injection 84
5.3. Illustrative results 85
5.3.1. Considered visualisation 85
5.3.2. Uniform noise applied to all topics from one sensor 86
5.3.3. Uniform noise applied to a single topic 86
5.3.4. Burst noise 91
5.3.5. Random message injection 93
5.3.6. Non-existent message injection 96
5.3.7. Conclusion 96
5.4. Experimental study of the ReCLiC parameters 98
5.4.1. The proposed noise scenario 98
5.4.2. Study of the internal aggregation operator 99
5.4.3. Study of the weighted average parameter 102
5.4.4. Study of the time window 104
5.5. Conclusions 108
6. Application to real data: trust dynamics of multiple sensors in the railway domain 109
6.1. Proposed visualisation of the trust evolution 109
6.1.1. Challenges of real data visualisation 109
6.1.2. Proposed heatmap visualisation 110
6.1.3. Trust temporal aggregation 110
6.1.4. Visualisation procedure 111
6.1.5. Section order 111
6.2. Result analysis 112
6.2.1. Global view 113
6.2.2. Global quality issues 115
6.2.3. Quality issue propagation effects 115
6.3. Summary 116
7. Conclusion and future works 119
7.1. Considered problem and challenges 119
7.2. Contributions 119

7.3. Future works 121

Bibliography 124

1. Introduction

According to the "Business Dictionary"[1], quality is a degree of excellence or a state of being free from defects or deficiencies. Quality problems constantly impact data and information, causing numerous problems with their further usage. One vivid and tragic example is the explosion of the space shuttle *Challenger* in 1986, where ten different categories of quality problems have later been identified as playing a role in the disaster, including inaccurate, incomplete and out-of-date data (Fisher & Kingma, 2001).

The first step in handling quality problems requires methods to identify and measure them. This is the objective of the *information scoring* task discussed among others by Batini and Scannapieco (2016): they aim at associating any pieces of information with a measure of their quality which can be numerical or descriptive, on a continuous or discrete scale. Information quality highly depends on the type of information, their content and considered domain. It is usually decomposed into different criteria, called dimensions, that allow to capture and combine different aspects of a piece of information. The huge variety of existing approaches, presented in Chapter 2, comes from differences in understanding, and thus defining, quality, its components and their combination, as well as the formal frameworks in which they can be represented.

1.1. Motivation and goals of the thesis

In this thesis, information quality is considered in the specific case where information is provided by sensors, i.e. devices measuring an aspect of reality and translating it into a digital value. Sensors can be common devices such as thermometers or weighing scales, as well as very specific devices dedicated to specialised domains, for instance, axle counters for railway signalling. The information produced by sets of sensors is often used to get enhanced knowledge about a given situation and the ability to differentiate between high and low information quality is a much-needed feature.

Indeed, sensors do not always provide correct information. There are many situations in which a sensor can fail, e.g. when encountering breakdowns, unfavourable operating conditions, communication problems or other interferences. There is no general framework or com-

[1] http://www.businessdictionary.com/definition/quality.html

mon model for assessing the quality of sensor outputs, although it may have many benefits such as comparing the quality of their output, providing consistency when scoring several types of sensors, simplifying their fusion and generally improving decision-aid systems.

Information scoring approaches dedicated to the information provided by sensors have been proposed, exploiting their specific property. However, as detailed in Chapter 2, they usually rely on external knowledge, in the form of ground truth or meta-information, which significantly limits their applicability in real cases. Indeed, ground truth is often difficult or expensive to acquire and meta-information is very dependent on specifics of the considered sensor.

The main goal of this thesis is to address these limitations by providing a model designed to be sensor-generic, not dependent on ground truth and dependent only on easy-to-access meta-information, exploiting only attributes shared among a majority of sensors. The proposed model is called ReCLiC from the four dimensions that it considers: Reliability, Competence, Likelihood and Credibility.

Informally, the ReCLiC model takes as input a log file of sensor entries and aims at attaching each log entry with numerical evaluation of its quality. Quality is understood here as the trust that can be put in the message content of the log entry, measured by considering the source, the content and the context of this message, which are the three main components defining a piece of information.

The research hypothesis is as follows: by applying a multidimensional information quality scoring model to the case of information provided by sensors, it is possible to obtain a relevant trust score for the evaluated piece of information using only easy-to-access meta-information and attributes common to all sensors.

1.2. Structure of the thesis

Chapter 2 presents the context of this thesis in more details, first discussing the notions of data and information. It then provides an overview of existing approaches for the tasks of data and information quality, detailing the specific case of sensors.

Chapter 3 introduces the proposed model, named ReCLiC, to score trust for information provided by sensors. It first details its requirements, to ensure the model is easy to apply and to use, independently of the type of sensor and their context and proposes to exploit common attributes, shared among most sensors. ReCLiC takes advantage of a general quality framework to define a model adapted to the case of information provided by sets of sensors. The chapter discusses in depth the requirements of the four proposed dimensions on which ReCLiC relies and proposes motivated definitions for each of them.

Chapter 4 goes from this generic ReCLiC model to a specific case: it describes the proposed implementation of the ReCLiC model to a real case, for a specific sensor in the railway signalling domain, called an axle counter. It discusses the form of the four dimensions for this case, lead-

ing to an operational instantiation of ReCLiC. In addition, it performs a formal study of the information scoring behaviour, analytically examining the effect of each dimension separately.

Chapter 5 experimentally investigates how the proposed ReCLiC model behaves: it proposes an experimental protocol based on realistic simulated data created from a real dataset containing log entries from sensors in the railway domain. This protocol allows to control various quality issues as well as their quantity, in four distinct scenarios of problematic log files, bypassing the absence of ground truth to evaluate the propositions. The chapter describes the conducted experimental study which includes a study of the parameters and allows to validate the proposed ReCLiC model.

Chapter 6 presents the application of the proposed ReCLiC implementation to analyse a real dataset where quality problems are detected and discussed. A new visualisation method is proposed to show multiple trust scores from many sensors at the same time. The chapter also proposes a dynamic point of view and in particular an analysis of trust propagation.

Chapter 7 draws the conclusion of the thesis and presents the perspectives it opens.

2. Foundations and key concepts

This chapter presents and discusses the key concepts manipulated in the thesis, providing a reminder about the necessary background knowledge as well as a brief review of state-of-the-art approaches. It first describes the domain of information scoring to which the contribution of this thesis belongs, defining it informally and discussing the main characteristics of existing methods. It then focuses on the specific task this work addresses, namely the case when the information to score is provided by sensors, characterising the task and presenting existing approaches. It finally evokes the question of such studies in a dynamic context, considering the temporal evolution of scores beyond the classical static cases.

2.1. Information scoring: measuring information quality

This section gives a broad view of the general task addressed by this thesis, namely *information scoring*. Intuitively, this task can be seen as enriching available information by a measure of quality: it makes it possible to distinguish between useless and worthy inputs, which is especially crucial for decision-aid systems.

Data and information quality evaluation has been analysed in domains of research like organisations management (Wang & Strong, 1996; Madnick et al., 2009), web information systems (Naumann, 2002) or information fusion (Rogova & Bossé, 2010; Todoran et al., 2013).

Information quality derives from data quality and this section discusses the differences between the notions of data and information and provides a short reminder about the topic of data quality. It then discusses the question of information quality, presenting the variety of existing approaches, it finally describes in more details the model proposed by Revault d'Allonnes and Lesot (2014), on which the contributions of this thesis rely.

2.1.1. Data versus information

We should first highlight that often, especially in common usage, both concepts of data and information are used with no clear distinction. An example is provided by the Cambridge Dictionary where *data* is defined as: „information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic

form that can be stored and used by a computer"[1]. We can see that *data* is basically defined as *information*, however pointing that it is more about fats or numbers. However, when looking up the meaning of *information* we can see that it corresponds to „facts about a situation, person, event, etc."[2]. None of these definitions allows us to understand the difference between these concepts.

Yet, the difference is stated by *Diffen* which is a tool that allows to compare similar concepts to see how they are different. It defines this subtle difference between *data* and *information* as: „**Data** are the facts or details from which **information** is derived. Individual pieces of data are rarely useful alone. For data to become information, data needs to be put into context."[3]. The individual definitions of both concepts state that „data is raw, unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized", and information is considered „when data is processed, organized, structured or presented in a given context so as to make it useful, it is called information".

These definitions make clear why we should consider them separately. Among others, because the information is based on analysis and interpretation of facts, it can be and it is often interpreted incorrectly, possibly leading to erroneous conclusions. It is crucial to be aware of whether, in a given situation, we are considering data or information to approach them accordingly.

Several authors in the literature provide definitions that are in agreement with the Diffen's ones. For Batini and Scannapieco (2016), the concept of data usually has a strong connection with measurements and numbers. Todoran et al. (2015) argue that data is context-independent, although data is collected with a clear purpose. Wang and Strong (1996) define information as an association between data and meaning. Thus, transforming data into information requires adding a context.

More generally, it is commonly considered that a piece of information consists of content, a source and a context. All three components are important and necessary to characterise a piece of information and are crucial for its scoring (see for instance Revault d'Allonnes and Lesot (2014)).

To illustrate the transformation of data into information, let us consider four numbers: $3, 15, 29, 7$. As such, these numbers have no meaning and are considered as data, they can be interpreted in many ways, e.g. seats in a theatre, a set of weights or a cypher to the safe. However, these numbers can be accompanied by knowledge about their source, e.g. a thermometer, and the context, temporal (e.g. 4 dates), spatial (e.g. indoor temperature of industrial cold storage) as well as regarding other sensors (e.g. thermometer of the neighbour room). They can then become information which can further be interpreted, for instance as cold, warm, normal or critical depending on the use context.

[1] https://dictionary.cambridge.org/dictionary/english/data
[2] https://dictionary.cambridge.org/dictionary/english/information
[3] https://www.diffen.com/difference/Data_ vs_ Information

2.1.2. Reminder about Data Quality

As discussed in the previous section, information and data are thus close, although, distinct notions: before considering the task of information scoring and information quality (IQ), it is useful to provide a reminder about the question of data quality (DQ). Most models decompose quality into different components, assessed by different dimensions called "quality criteria". In the beginning, the quality notion was limited to accuracy and currency. Since then, abundant literature has been developed, proposing long lists of other crucial criteria. A survey compiling over 40 of them is presented by Sidi et al. (2012). This section presents in more details two reference models that propose structured views of dimensions that can be considered for evaluating data quality.

Wang and Strong (1996) model Wang and Strong (1996) were among the first arguing that limiting quality to the level of accuracy is not enough. They highlight that the level of quality for given data can depend on its purpose. In order to better measure quality, they conducted a two-stage survey among data consumers and a two-phase sorting study to develop a hierarchical framework for organising data quality dimensions. As a result, they propose four categories, detailed below, which should be addressed when managing data quality. This model has been successfully used in industry and government bureaucracy showing the importance of a multidimensional approach (Wang & Strong, 1996). This work does not distinguish between the notions of data and information: even though they use the term "Data Quality" in the proposed categories, they are in fact considering both DQ and IQ.

More precisely, Wang and Strong (1996) propose 15 dimensions grouped in 4 main categories by analysing and grouping 179 different attributes. We reproduce them below:

1. Intrinsic DQ refers to the source and the content.
   (a) Accuracy - data are certified error-free, correct, flawless, errors can be easily identified, the integrity of the data, precise
   (b) Believability
   (c) Objectivity - unbiased
   (d) Reputation - a reputation of the data source, a reputation of the data
2. Contextual DQ highlights that data quality needs to be evaluated with the considered task in mind and its context.
   (a) Relevancy - applicable, interesting, usable
   (b) Completeness - breadth, depth, and scope of data
   (c) Timeliness - age of data
   (d) Value-added - data give you a competitive edge, data add value to your operations

   (e) Appropriate amount of data

3. Representational DQ requires the data to be easy to understand and interpret, among others meaning that the data is presented in a clear manner.

   (a) Interpretability

   (b) Ease of understanding - clear, readable

   (c) Representational consistency - data are continuously presented in the same format, consistently formatted, data are compatible with previous data

   (d) Concise representation - well-presented, concise, well-organised, aesthetically pleasing, a form of presentation, well-formatted, format of the data

4. Accessibility DQ highlights the importance of the data system to be obtainable by the data consumer.

   (a) Accessibility - retrievable, speed of access, available, up-to-date

   (b) Access security - data cannot be accessed by competitors, data are of a proprietary nature, access to data can be restricted, secure

These groups and dimensions represent a holistic approach to define and measure data quality. In a specific scenario, not all of them are required or even possible to measure, however it allows for a basic classification of all existing criteria.

Batini and Scannapieco (2016) model Similar conclusions are presented by Batini and Scannapieco (2016) who propose a classification framework where dimensions are included in the same cluster according to the similarity of the characteristics they measure. They end up with eight categories named after their representative dimension, which we reproduce here:

1. **Accuracy**, correctness, validity and precision focus on the adherence to a given reality of interest.

2. **Completeness**, pertinence and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest.

3. **Redundancy**, minimality, compactness and conciseness refer to the capability of representing the aspects of the reality of interest with the minimal use of informative resources.

4. **Readability**, comprehensibility, clarity and simplicity refer to ease of understanding and fruition of data by users.

5. **Accessibility** and availability are related to the ability of the user to access information from his or her culture, physical status/functions and technologies available.

6. **Consistency**, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms.

7. **Usefulness**, related to the advantage the user gains from the use of information.

8. **Trust**, including believability, reliability and reputation, catching how much information derives from an authoritative source. The trust cluster encompasses also issues related to security.

Comparison As we can see, there are similarities as well as differences between the two above classifications. For instance, whereas in Wang and Strong (1996) believability and reputation are considered as intrinsic dimensions, alongside accuracy, in Batini and Scannapieco (2016) there are considered as part of trust evaluation. However, most of the dimensions have a clear understanding of what they represent.

These are two examples of quality criteria surveys, others exist as well, see e.g. Sidi et al. (2012): they differ by the list of considered dimensions and their organisation. As illustrated by the two detailed cases, no consensus can be found in the literature regarding these quality criteria.

Even though these authors mostly discuss data quality, the difference with information quality is sometimes small and some criteria can be considered for both notions. Furthermore, since information quality uses content, among other components, data quality evaluation is usually necessary.

2.1.3. Information quality

As mentioned previously, a piece of information consists of at least three main components: source, content and context. Because of that, scoring information allows for a more thorough evaluation, than the criteria involved in data quality. However, the same principle usually applies: the notion of quality is decomposed into different dimensions that allow to capture and combine different aspects, in a multi-criteria aggregation framework.

Existing models differ by the considered dimensions and their aggregation, as well as by the formal frameworks chosen to measure the criteria and perform their combination. It should be underlined that, in addition, authors often do not agree on the name used to describe a dimension or, on the contrary, on the definition of a dimension: two different names can refer to the same dimensions, and a given name can refer to two different notions. This makes the comparison between existing models difficult. Moreover, due to the context of the considered issue and user's needs, two authors can choose different dimensions to describe quality for seemingly the same problem.

This section illustrates the variety of IQ dimensions as well as some considered formal framework used to perform information scoring.

IQ dimensions Some DQ dimensions can be also considered for IQ. However, since information quality consists of multiple components, new dimensions are also considered that can be strictly related to IQ rather then DQ. We list some of them below, often observed in the literature, as an example of their variety:

- Verifiability - is the degree and ease with which the information can be checked for correctness, especially if the information is mistrusted (Naumann, 2002)
- Reliability - relative stability of information content considered from one time to another under the same environmental characteristics, e.g. sensor readings under the same conditions (Rogova & Bossé, 2010); an evaluation of the source independent of all information (Besombes & Revault d'Allonnes, 2008)
- Relevance - providing useful information regarding a given question of interest (Pichon et al., 2012)
- Truthfulness - actually supplying the information the source possesses in its full extend (Pichon et al., 2012)
- Proficiency/competence - the ability of the source to provide useful information in a given situation, it is topic-dependent and thus varies from one piece of information to the other (Besombes & Revault d'Allonnes, 2008; Pichon et al., 2014; Revault d'Allonnes, 2014; Lesot & Revault d'Allonnes, 2017; Lenart et al., 2018)
- Likelihood - qualifying an information based on the user's global take on the state of the world (Besombes & Revault d'Allonnes, 2008; Lenart et al., 2018)
- Credibility - a degree of confirmation resulting from comparing of the piece of information to be rated with other available information (Appriou, 1998; Pon & Cárdenas, 2005; Florea & Bossé, 2009; Florea et al., 2010; Rogova et al., 2013; Lesot & Revault d'Allonnes, 2017; Lenart, 2018)
- Sincerity - the tendency of the source to tell the truth or not (Lesot & Revault d'Allonnes, 2017)
- Plausibility - the compatibility of the considered piece of information with the rater's knowledge and his opinion (Lesot & Revault d'Allonnes, 2017)

As can be seen from the above definitions, for most information quality dimensions, it is necessary to obtain additional knowledge and they cannot be assessed based only on a piece of information itself. For instance, likelihood requires to consider real-world properties in a given context, similarly, relevance is strongly connected with the considered question of interest.

Formal frameworks Existing models also vary regarding the formal framework they consider to measure the values of the dimensions and to represent the final quality score. Among them we can recognised classical mathematical theories like the probability theory, the fuzzy set theory (Zadeh, 1965), the possibility theory (Dubois & Prade, 1988; Lesot et al., 2011), the Dempster-Shafer theory (Shafer, 1976; Appriou, 2001; Bovee et al., 2003; Samet et al., 2014), as well as more recent ones, like the generalised information theory (Klir, 2005), that constitutes a variation of probability and possibility theories, or the multivalued logic framework (Revault d'Allonnes & Lesot, 2014).

In our research, we consider a simple multi-criteria aggregation framework, where each criterion is measured on a $[0,1]$ scale, as well as their aggregation to the global quality score.

2.1.4. Focus on a trust model

One family of information quality models focus on trust, that measures the level of confidence put in a piece of information. Trust is also viewed as a composite criterion, existing models for assessing trust vary in the dimensions used for its representation as well as in the understanding of the concept it measures. For instance, Young and Palmer (2007) define trustworthiness as the degree to which a piece of information (from a source) is considered conforming to fact and therefore worthy of belief; Trust is a notion that is also used in other contexts than information scoring: it can, for instance, apply between agents, as proposed by Demolombe (2004) who defines trust as the mental attitude of an agent towards another agent where an agent can be a human, a robot, a sensor or a program. Trust is also widely used in the context of the Internet of Thing, see e.g. Baqa et al. (2018).

This section details the trust model for information scoring proposed by Revault d'Allonnes and Lesot (2014), on which the contributions of this thesis rely. Their approach does not focus on any particular case and can be applied for any type of considered information: they propose a general framework that considers four dimensions to update dynamically the evaluated trust value. These four dimensions aim to answer different questions, starting from a general one that only depends on the source, not taking into account the content of the piece of information, to the most specific, dependent on the content and the context.

The four dimensions are chosen to answer the following questions: 'Who is telling me?' (reliability), 'What does he know about the topic?' (competence), 'How likely does it seem?' (plausibility), 'Is anyone else reporting it?' (credibility). The first two define the source, reliability is constant, independent on the current information it produced, where competence depends both on the source and the piece of information. The second two relate to the information content, plausibility measures the compatibility of information with the rater's knowledge and credibility takes a degree of confirmation of the considered piece of information from other sources. To measure these dimensions, the authors propose to use an extended multivalued logic framework.

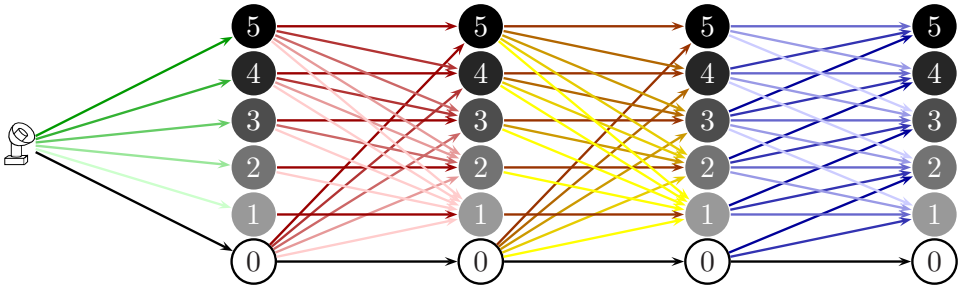


Figure 2.1: Sequential projection of the trust building process, from Revault d'Allonnes and Lesot (2014)

To combine the selected dimensions, the authors propose first to order them from the most general to the most specific, both content and contextwise. Because of that, the evaluation starts with the global dimension and it is progressively corrected with more specific characteristics. For the updates, an abating operator is used, except for the last step. By starting with reliability as the most general dimension, updating the first score with competence can result in either lower value or, at most, the same one. Indeed, if the source is unreliable, it does not matter if it is competent in a given topic and the current trust score cannot be increased. A similar situation is considered when updating trust with the following dimension: plausibility. Only for credibility, the second type of operator is considered. Since this dimension is based on confirmation from other sources, it can both increase or decrease the current trust score.

These two types of operators are illustrated in Figure 2.1, p. 28 where arrows represent the update operator of each dimension. We can see that arrows from each score point downwards except in the last one that represents credibility. In addition, the authors include the special value denoted as 0 which is used when the dimension cannot be evaluated. The authors highlight that if the evaluation of trust was possible at any stage, it is also possible in all subsequent stages.

In Chapter 3 this approach is adapted to the case of sensors (see the next section) by redefining the dimensions, proposing an approach to evaluate them as well as new ways to aggregate them.

2.2. Information quality for sensor measurements

The previous section discussed general approaches to information and its quality, here we focus on the specific case considered in this thesis, namely information provided by sensors. The reason for focusing on this type of information is to exploit their specificity to define a dedicated scoring method.

2.2.1. Sensor definition and characteristics

A sensor is a device which translates a part of reality into a digital value. Sensors can be common devices such as thermometers or weighing scales, as well as very specific devices dedicated to specialised domains, for instance, axle counters for railway signalling. One can distinguish two major types of sensors: asynchronous event-based ones and continuous ones. The first one produces messages when an event happens e.g. the light is turned on or the temperature reaches 30°. It is not known beforehand when the event will happen, thus the messages are produced irregularly. The second type groups sensors that continuously provide information about a given situation, e.g. a thermometer which measures the current temperature every few seconds. Here, messages are produced in regular intervals and problems can be assumed if it is not the case. Of course, there can be some cases that mix both approaches, e.g. a thermometer that reports temperature only when it changes, however, such cases are not common.

Information coming from any device might be subject to imperfections, caused by mechanical breakdowns or information flow disruptions. With the growing utilisation of sensor networks, the potential cost of errors provoked by anomalous sensor responses is becoming increasingly important. These errors can influence the information the sensor produces in a way that is not easy to recognise and understand. Sensor systems are exposed to multiple operational risks and because of that, sensor systems vulnerability has been examined in many fields like automated vehicles (Petit & Shladover, 2014), a global positioning system (GPS) (Grant et al., 2009), a maritime navigation devices (Balduzzi et al., 2014; Iphar et al., 2015; Laso et al., 2017) to name a few. A particular effort is necessary to ensure high-quality information in cyber-physical systems in order to eliminate anomalies and automatically evaluate the relevance of sensor information streams and deliver this information to decision-aid systems.

Another characteristic common to the majority of sensors is that they usually are able to monitor their behaviour and inform about some problems by sending predefined error messages. This raises the question of whether these messages also should be subject to information quality evaluation. Since these messages highlight any abnormal behaviour, it is important to address the potential problem even if the error message is scored low quality. An answer to this question is proposed later in the thesis.

Of course, it can be observed that sensors often produce information on more than one topic. This is, for instance, the case with complex sensors or when one type of sensor is used in different contexts. For instance, when we take two thermometers and locate one inside a house and the other outside the house, we know that their readings are not comparable and they are treated as providing information about different topics. Another example is an object recognition system which recognises between planes and helicopters in different accuracies. These two categories are treated as two topics.

An example of pieces of information provided by a sensor is presented in Table 2.1. We consider that information provided by sensors is stored in a *log file* or *database*. Each row is

Table 2.1: Example of input data structure

id	Date	Time	Sensor ID	Topic	Message
1	11.03.2015	07:24:53	S1	T1	m1
2	11.03.2015	07:25:40	S1	T2	m2
3	11.03.2015	08:23:18	S1	T3	m1
4	11.03.2015	08:24:08	S2	T5	m2
5	11.03.2015	09:15:23	S2	T4	m2
6	11.03.2015	09:16:08	S1	T1	m2
7	11.03.2015	09:39:45	S3	T7	m3
8	11.03.2015	09:40:29	S3	T8	m1
9	11.03.2015	10:22:14	S3	T7	m2
10	11.03.2015	10:23:03	S4	T9	m1

called *log entry* and consists of many attributes e.g. date and time, the id of the sensor or topic of information. The provided value, either numeric or descriptive, is called *message*. In this example, there are four sensors altogether covering eight different topics and producing three different types of messages.

2.2.2. Existing models

There is no general framework or a common model for assessing the quality of sensor outputs, although it may have many benefits such as comparing the quality of their outputs, providing consistency when scoring several types of sensors, simplifying their fusion and generally improving decision-aid systems.

This section first discusses the two most generic approaches found in the literature. Most other approaches (Pon & Cárdenas, 2005; Guo et al., 2006; Florea & Bossé, 2009; Destercke et al., 2013) transpose generic information scoring models to the case of sensors, considering specific dimensions to combine. This section details, in turn, the two main dimensions commonly used, namely reliability and credibility, for which many definitions have been proposed.

2.2.2.1. Generic models

Bisdikian et al. (2009) propose a generic principle for creating a system based on sensors: by including additional meta-information, it allows to improve the information quality evaluation at the production level. They propose a generic form, that can be applied in various contexts for this meta-information. The main difficulty with this proposition is that it only can be applied for creating new sensor systems and there it is not possible to use it with existing ones.

Todoran et al. (2013) consider the fusion of information from many sensors to improve decision-aid systems. They differentiate between the local IQ which considers a single source with its messages and the global quality of the entire system based on dynamic local sources which can be dynamically added and removed. Their work is also based on integrating meta-information with local sensor sources to highlight how they impact the global system. This

work considers a global framework, but it does not propose how to evaluate IQ from local sensors, which is often a difficult task.

2.2.2.2. Reliability as the main quality criterion

Reliability is generally understood as the ability of a system to perform its required functions under some stated conditions for a specified time, as in the works of Blasch (2008); Florea et al. (2010); Destercke et al. (2013). Assessing reliability is not easy as it often requires information difficult to acquire, as detailed in the following.

Florea and Bossé (2009) as well as Destercke et al. (2013) consider using meta-information of the source, e.g. its specification, protocol or environment. The gathered knowledge is then combined to propose the reliability score. This approach is limited to the case where valuable meta-data are available and it is limited to the considered scenario, making it very specific and difficult to generalise.

Another common approach to define reliability consists in viewing it as accuracy, in the case where ground truth is available, i.e. knowledge about expected results (Blasch, 2008; Florea et al., 2010; Martin et al., 2014). In this case, accuracy is the proportion of cases where the provided information indeed describes reality. This definition relies on the assumption that, if a source was accurate in the past, it will be accurate in the future. Now, this hypothesis is a strong one: a previously accurate source may suddenly become inaccurate in the future, e.g. due to a failure, something that is not considered in this case. Moreover, this definition suffers the limitation that it requires ground truth to be available. However, ground truth is not always available or it is expensive to acquire.

Mercier et al. (2008) propose to score reliability such that it better reflects the reality of the sensor and its working environment by enriching it with its context. Then, different situations can result in different output qualities for a given sensor. For instance, in the case of target recognition (Mercier et al., 2008), the performances of a data acquisition system may depend on background and target properties, making the reliability of the decision system dependent on the target at hand. A sensor that recognises between three objects (e.g. helicopter, aeroplane and rocket) can have different accuracies for each one, effectively creating a vector of three reliabilities with different contexts.

It has also been proposed to consider accuracy as a component of a reliability score that can be enriched with other components. For instance, Blasch (2008) enriches the previous definition by considering that reliability requires accurate and on-time results. However these are not always achievable simultaneously: sometimes having more accurate data leads to longer collecting time, which induces a trade-off between accuracy and timeliness.

2.2.2.3. Credibility and its role in scoring information

Credibility is generally defined as the level of confirmation of a given piece of information by other, independent, sources and constitutes another classical component of information

quality reference needed both regarding the definition of credibility and its common use. Relying on external sources differentiates credibility from most other dimensions which focus on the evaluated piece of information. It is often evaluated as the level of consensus among a group of sources with the considered one, which means that all these sources should provide the same information on the same topic for credibility to be high (see Section 2.1.4). It raises the question of aggregation or a fusion of information provided by other sources.

Credibility is often considered in the case of sensors: usually in any kind of production systems, there are many devices that interact with each other and make it confirm or contradict each other. Many different approaches have been proposed, depending on the considered devices and context, some of them are described in the following.

Appriou (1998) uses credibility to score the quality of the evaluated piece of information. He argues that, often, measurements are more or less of doubtful origin and prior knowledge about the sensor is poorly defined. In that situation, any quality criterion based strictly on a piece of information or its source is not efficient. Credibility can then be a key to score quality, as it does not depend on the 'unknown' reliability of the source. This makes scoring quality external, independent from any knowledge about the source, except the available other sources. In the example of the target tracking process he proposes, Appriou (1998) aggregates the output from multiple devices to improve the results: when a sensor informs about the presence of an object in the area, this piece of information is confronted with others. If enough sources confirm the information, the result gets higher confidence, otherwise, the information is treated as a false positive. This is the simplest approach of using credibility in the case of sensors, it is based on the number of confirming messages. Similar approaches are proposed by many authors. For instance, Xiao et al. (2010) consider the collaboration of multiple ultrasonic sensors which are used to improve information about target positions. Hermans et al. (2009) analyse results from multiple sensors to estimate the quality of sensor data providing better quality information for decision-aid systems. Zahedi et al. (2008) use output from multiple sensors to attach each sensor with the score of its normal/faulty behaviour to be later used in information fusion.

Guo et al. (2006) propose to incorporate unsupervised learning in order to monitor the degradation of a sensor in time in an automatic way using a credibility notion. Their proposition is to check if the results from one sensor are consistent with the results from other sources when providing information on the same topic. This principle is the base of an automatic system that continuously monitors a potential degradation or a failure of a sensor and alarms the user if messages from the considered sensor are not consistent with the others. This can ensure high-quality data that will not harm the decision making or the performance of the system. Their approach is similar to the previous propositions and credibility, which they name *dynamic reliability*, is evaluated by scoring a degree of consensus among a group of sensors.

In the work presented by Shao et al. (2017) credibility and cooperation are considered in the area of wireless sensor networks. They rely on the hypothesis of a spatial correlation be-

tween sensors where the distance between sources is important in credibility scoring. In order to decrease data transfers and energy usage, a sensor with suspicious activity sends an error diagnosis request to its close neighbours which in turn send their diagnoses. A suspicious sensor then decides its status based on these replies. The result classifies this suspicious activity either as an error or a rare message, normally not seen in this context. The other sources on which credibility is based are in this case neighbours which can confirm a potential issue with the evaluated sensor.

There also exist more complex systems that use credibility on multiple stages. Florea et al. (2010) propose a framework that consists of two steps. First, the sensors are grouped so that each group contains devices that provide information about the same topic and can easily be compared. The similar sensor fusion (SSF) function is introduced to combine values from each group, mostly based on a simple consensus between sensors, as is the case in the approach described by Guo et al. (2006). Next, the representative values from all groups are combined with a dissimilar sensor fusion (DSF) function. It is more complex and needs to include specific attributes of each group and how they relate one to another. The aggregation rules of SSF and DSF are not always the same and depend on the sensor types and user needs. The DSF step introduces additional connections and complexity as two values representing different groups are not comparable since they represent different topics.

One of the important aspects of credibility is the number of other sources used for its scoring. Chen et al. (2016) argue that more sources do not always mean a better outcome. In their work, they consider a surveillance system that analyses how the number of sensors influence the number of false alarms and missed alarms. To do it, they provide an analysis of how this number changes depending on the total number of sensors. They show that too many sensors increase the number of false alarms significantly, whereas too few lead to a great number of missed alarms. They also propose a statistical method to find the optimal number of sensors.

It can be observed that in most of the literature, credibility is defined by considering the level of confirmation by other sources: this approach implicitly implies that if a message does not confirm the information then it contradicts it. This is a major drawback as this is not always the case, there is a possibility that the considered message can be neutral towards the evaluated one and treating it as contradiction can lead to false results. This is one of the problems that we address in this thesis.

2.3. Dynamic analysis: quality evolution

Most works described in the previous section deal with the evaluation of a single piece of information. However, sensors usually produce many pieces of information during the considered period of time: analysing how these trust values change can be crucial to interpreting their behaviour. In the literature, this topic is rarely considered and especially in the case of sensors.

However, there are few studies on the general issue of trust evolution. They examine the possible successive values trust can take: trust depends on the information flow, its context and the way previous pieces of information have been scored (Jonker & Treur, 1999; Falcone & Castelfranchi, 2004; Cvrcek, 2004; Lenart et al., 2019).

Jonker and Treur (1999) argue that a new piece of information that influences the degree of trust is either trust-positive, i.e. it increases trust to some degree or trust-negative where trust is decreased to some degree. The degree to which trust is changed differs depending on the used model. They distinguish between two types of trust dynamical models: a trust evolution function considering all previous information and a trust update function storing only the current trust level and having the ability to include the next information. Both considered in this thesis.

Mui (2002) proposes an asymmetrical approach for the increase/decrease trust rate, which is inspired by the approach observed in humans: trust increases slowly but decays rapidly. This asymmetrical behaviour is also arguably important from a practical point of view, for many applications (see e.g. the case of malicious attacks in the security domain (Duma et al., 2005)).

The problem with these approaches is that they more appear towards assessing trust for the source, and indirectly, for a piece of information it produces, rather than focusing on the piece of information itself. Moreover, the authors often consider only human as a possible source of trust changes. In our research, we propose to reverse the dependency and by assessing trust for individual pieces of information, we want to study possible behaviours of the source.

2.4. Summary

In this chapter, we discussed the differences between data and information and different approaches to scoring information quality, both in general and in the case of sensors. We showed that there are numerous possibilities, with their advantages and disadvantages, and there is a lack of a single general approach that can suit different scenarios of sensor information.

The main goal of this thesis is to address these limitations by providing a model designed to be sensor-generic, independent of ground truth and dependent only on easy-to-access meta-information, exploiting only attributes shared among the majority of sensors.

Moreover, we aim to address trust evolution and study its behaviour when encountering different problems. The current approaches in the literature are mostly limited to human sources and are largely limited to scoring evolution of trust for the source, rather than the message itself.

3. Dynamic trust scoring for sensors: the proposed ReCLiC model

This chapter describes the model we propose to assess the quality of information provided by sensors, named ReCLiC after the four components it integrates: Reliability, Competence, Likelihood and Credibility, where quality is understood as the trust that can be put in the considered information. After detailing the goals and technical requirements regarding the inputs it relies on, the chapter provides an overview of the ReCLiC model and then describes, in turn, each of its four components, as well as their final aggregation into a trust value.

A preliminary version of the ReCLiC model have been proposed in (Lenart et al., 2018; Lenart et al., 2019) and specifically regarding credibility in (Lenart, 2018), they also have intersections with the works presented in Chapters 4, 5 and 6.

3.1. Goals and requirements of the proposed model

Informally, the ReCLiC model takes as input a log file of sensor entries and aims at attaching each log entry with a numerical evaluation of the quality of this entry, measured by considering the source, the content and the context of the entry, which are the three main components defining a piece of information (see Section 2.1).

As discussed in the previous chapter, existing models for measuring the quality of the information provided by sensors make the assumption that ground truth or meta-information is available, limiting their genericity. This section states the assumption on which the proposed ReCLiC model relies on, first describing its desired characteristics and then precisely presenting the inputs it requires.

3.1.1. Desired characteristics

Sensor-genericity Huge attention is placed for ReCLiC to be able to apply to all cases where information is provided by sensors. Most of the approaches discussed in Section 2.2, p. 28, rely on adapting to the situation at hand, by incorporating its specific properties. Such dedicated approaches offer advantages but they lack genericity. In our model, we limit the assumption of

the provided attributes to only the ones that are shared among a majority of sensors. We discuss them in more detail in the next section.

Non-dependence to the ground truth Ground truth refers to information about the reality measured by the evaluated sensor. If one has access to complete knowledge of the situation that the sensor measures, the provided message can be confronted and the accuracy of the device evaluated. Most approaches rely on ground truth to train a model on the specific data, that can later be used to assess the quality level of the piece of information provided by the considered sensor (see Chapter 2). However, ground truth is often unavailable or expensive to obtain. It is often the case with sensors where their readings may not be possible to obtain by other means. In these cases, methods that rely on ground truth cannot be applied.

Limited meta-information dependency Another problematic aspect refers to meta-information which is sometimes considered. Meta-information is understood as any additional information to the one provided by the considered device. This type of information can rarely be generalised and a method that considers very specific meta-information is often limited to be only used in one specific case. In our model, we decrease the reliance on this type of information to a minimum. ReCLiC only considers meta-information that is easy-to-access and applies to the majority of sensors. Moreover, if necessary, it can be derived from the data itself with limited compromises, as detailed in the following sections. This approach is studied and illustrated in a real case from the railway signalling domain.

3.1.2. Considered data input and notation

In order to achieve the characteristics described in the previous section, the ReCLiC model does not impose restrictive constraints on the data to use: it only relies on a log file containing common attributes, as well as some easy-to-access meta-information as described in Section 3.1.3.

Regarding the available inputs, ReCLiC relies on the very general log file schema mentioned in Section 2.2.1, p. 29, and illustrated by the schematic view given in Table 2.1, p. 30: each message comes along with a date, a source ID and a topic that altogether represent a single log entry. It is important to highlight that all these attributes are common for any sensor as every message comes with a timestamp and includes an ID of the device that reported it. Only the topic may not always be present: it can then be considered that each sensor produces information regarding a single topic.

The aim of this thesis is to add a column in Table 2.1, p. 30, with the quality value associated to each piece of information. It is a numerical value that represents the trust that can be put in the message content of the log entry. This trust is an aggregated value of four different quality

criteria that incorporate not only the message content assessment but also include its source and context.

Formally, throughout the thesis, the following notations are used: $\mathscr{L}$ denotes the complete set of log entries in a given database, $\mathscr{L}_s$ is the set of log entries produced by sensor s and $\mathscr{L}_{s,z}$ is the set of log entries produced by the considered sensor s on topic z. The notation l corresponds to one log entry defined as four values: *l.fullDate* corresponding to date and time, *l.sensor* to the sensor ID, *l.topic* to the topic of the message and *l.message* to the provided piece of information describing the sensor's state. The set of all sensors is denoted by $\mathscr{S}$, the set of all timestamps by $\mathscr{T}$ and the set of all topics by $\mathscr{Z}$.

An example of these notations can be illustrated based on Table 2.1. $\mathscr{L}$ contains all log entries, i.e. $|\mathscr{L}| = 10$. Messages from a given sensor depend on column "Sensor ID", e.g. $\mathscr{L}_{S1} = \{1,2,3,6\}$ or $\mathscr{L}_{S2} = \{4,5\}$. When considering both sensor and topic, one for instance has $\mathscr{L}_{S1,T1} = \{1,6\}$, $\mathscr{L}_{S1,T3} = \{3\}$ and $\mathscr{L}_{S3,T7} = \{7,9\}$. The set of all sensors is $\mathscr{S} = \{S1, S2, S3, S4\}$, the set of all topics: $\mathscr{Z} = \{T1, T2, T3, T4, T5, T7, T8, T9\}$ and the set of all timestamps contains date& time of all log entries.

3.1.3. Meta-information input

As discussed in the previous section and Section 2.2, p.28, meta-information is often the weak point of any proposed method by limiting the approach to its specific case. However, some features are common to many sensors thus making it possible to use them in the scoring process.

ReCLiC proposes to use two types of meta-information that describe the work processes of the considered devices, respectively detailed in turn in this section: the state transition graph and the sensor network. Both can be considered as easy-to-access and automatically extracted from the data themselves if necessary, as is the case for the real data we use (see Chapter 4).

3.1.3.1. State transition graph

Definition and usage The *state transition graph* represents the admissible consecutive states for the considered device when the messages it produces are considered as indicating its state. The last produced message of a sensor is considered as its current state, state changes happen when new messages are produced. The purpose of the graph is to represent valid transitions between two consecutive messages: an edge between two nodes indicates that it is possible to have these two subsequent messages.

An example of a general state transition graph is presented in Figure 3.1 where a sensor can produce five different messages, which in this case represents five different states. Transitions show how the sensor is expected to work. For instance, when the current message is *State1* then the following one can only be *State2* or *State3*, any other one (including *State1* itself) is not possible.

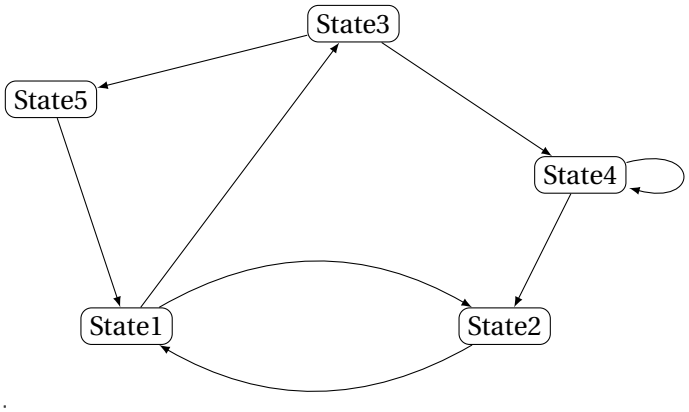


Figure 3.1: Example of a state transition graph

In subsection 2.2.1, p. 29, we discussed two types of sensors: asynchronous event-based ones and continuous ones. A state transition graph can be created for both. For event-type sensors, different messages represent nodes in the graph. For continuous ones, if the message is a numerical value, e.g. a thermometer producing temperature in degrees as an output, defining the states is somewhat flexible: each state can be a separate temperature value or a range of values. A transition between two states means that it is possible to have this quick change in temperature in a given time. For instance, it is possible to have two consecutive readings that show 2° after 1° but it shouldn't be possible to have 10° right after 0°.

Acquisition There are two possibilities to acquire a state transition graph for a considered device: the first one relies on meta-information e.g. provided by an expert or the system specification. Creating a graph out of this knowledge is straightforward.

The other possibility is to analyse data produced by sensors of the same type: nodes can then be created by searching all possible types of produced messages. Second, historical data can be processed such that each two consecutive log entries are considered as an admissible transition between two states in the graph.

This data-driven approach has the obvious advantage over the expert one that it does not rely on external input. However, it can lead to two types of errors: erroneous edge absence or erroneous edge presence. The first case is caused by some transitions which are rare and appear only in special cases. In the part of data used for creating the state transition graph, these transitions might be absent, thus the corresponding edges are not created. This can lead to false problem recognition when using the graph later on. In the second case, the part of data used for creating the graph can be corrupted and include a transition that in reality does not exist. This will result in a graph that will not recognise this issue in future usage. To mitigate this problem, a threshold can be introduced to limit possible transitions to the ones that are present in the data from multiple sensors or extracted in a sufficient amount. However, this solution is not ideal and might lead to discarding valid transitions.

Both approaches have their respective advantages and disadvantages. Due to the requirements stated in subsection 3.1.1, we favour the second, data-driven one. When possible, we consider its combination with the expert one. In such a case, analysing the possible differences between the two obtained graphs. It allows additional insights contributing to the study of the database quality (see subsection 4.2.3.1, p. 68).

3.1.3.2. Network of sensors

The *network of sensors* is also a graph, whose nodes represent the sensors and an edge between two nodes indicates that these sensors should be correlated. Each edge can be weighted, representing the level of this expected correlation, however, in practice, it is often difficult to compute correct weights. The correlation can be connected with the concept of spatial confirmation meaning that physical geographical proximity between two sensors provides some knowledge that can be used in the scoring process. The activity of one sensor which is related to the other means that one can expect their messages to be correlated as well, for instance, an event that triggers the first source, also impacts the second one, resulting in messages from both sources.

This approach is similar to the affinity graph considered in the information scoring model proposed by Lesot et al. (2011) which also groups sources according to the expected correlation, in their case, whether two sources manifest affinity or hostility relations.

The method we propose to acquire the sensor network is described, with its exploitation, in subsection 3.6.2, p. 50.

3.2. Overview of the proposed approach

This section provides an overview of the ReCLiC model we propose, presenting its general principle and offering a graphical representation.

3.2.1. General principle

As is often the case for information scoring (see Chapter 2), the information scoring model we propose consists of aggregating several dimensions, that capture and combine different characteristics of the current information to be scored. The proposed ReCLiC model is named after the four components it integrates, namely Reliability, Competence, Likelihood and Credibility.

These chosen dimensions are inspired by the general model proposed by Revault d'Allonnes and Lesot (2014), see subsection 2.1.4, p. 27. As in their case, ReCLiC also starts from the most general criterion which provides an a priori evaluation of the source, in the form of reliability. However, it is necessary to adapt this notion to the case of sensors. In Section 3.3 we propose to define this a priori evaluation as an assessment of whether the sensor appears

to work properly at a given date. Note that in the model proposed by Revault d'Allonnes and Lesot (2014), reliability is not explicitly a dynamic notion, that depends on the date, whereas this seems to be a crucial property in the case where the information is provided by sensors.

In a second step, the source is evaluated in the respect to the content it provides, by scoring competence. This second source evaluation is less general, it evaluates whether the source can provide information about a given topic in a given time. As for reliability, we make this notion a dynamic one. In Section 3.4 we adapt this criterion to the case of sensors, it is similar to reliability but allows to differentiate the ability of a sensor to provide quality messages in different topics.

In the model presented by Revault d'Allonnes and Lesot (2014) the next dimension, named plausibility, checks if the message is compatible with what the rater knows, i.e. his background knowledge. This may be advocate for the need of external knowledge, which can often mean that a ground truth is necessary. As it would invalidate our desired characteristics of the ReCLiC model, we propose to replace plausibility with likelihood, described in Section 3.5: instead of relying on external knowledge, it focuses on the internal aspect of information flow and it answers the question: „does the message content seem appropriate with respect to the previous messages?". Alternatively, this definition can be understood as a question of temporal credibility which checks for corroboration from the past.

The final dimension matches the one from Revault d'Allonnes and Lesot (2014) and most other models. It is an essential context checking step represented by credibility. In general, it is used to search for confirmation from other sources. In ReCLiC we present an approach for the sensor case in Section 3.6, p. 49. We incorporate an idea of geographical corroboration that exists between devices that are close to each other to choose which sensors and what messages can be used as a confirmation or invalidation.

The proposed dimensions are aggregated into a single trust value. This value describes the level of confidence for the message produced by the sensor in a given context. The way of aggregating the considered dimensions differs from the approach presented by Revault d'Allonnes and Lesot (2014): whereas in their work, dimensions are combined in a specific order that aims at updating trust with progressively more refined characteristics of the considered piece of information (see Section 2.1.4, p. 27), in the ReCLiC model the aggregation function is proposed but it can be modified, which allows to change the desired behaviour. In addition the approach presented by Revault d'Allonnes and Lesot (2014) gives a result in a discrete scale of six values, in an extended multivalued framework. In the ReCLiC model, the final value, as well as each dimension, are set in the continuous scale in $[0, 1]$ range.

Major differences between models The three major differences between the ReCLiC method and the one presented by Revault d'Allonnes and Lesot (2014) are listed below.

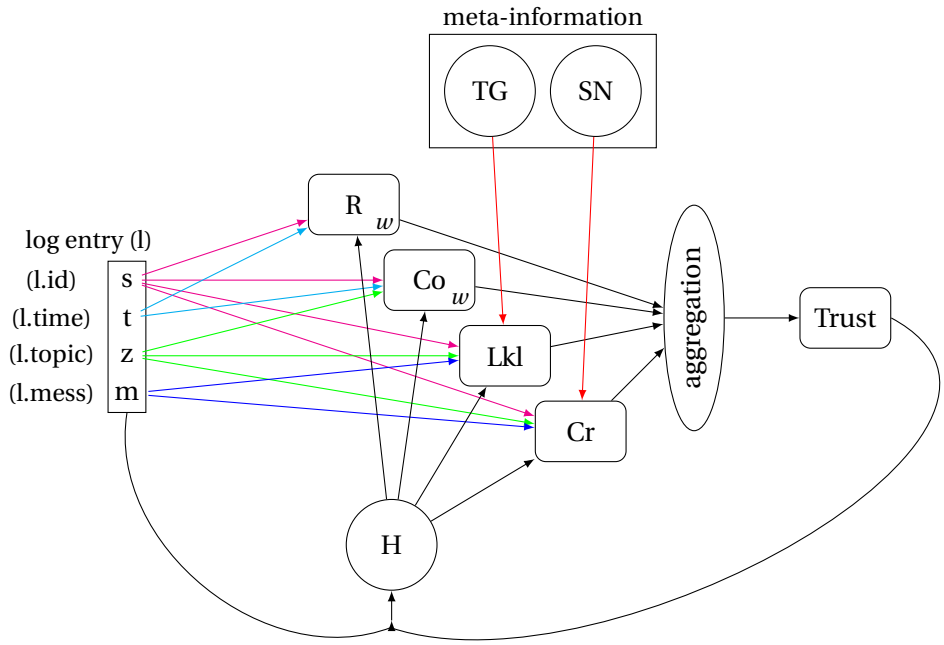


Figure 3.2: The general ReCLiC model.

1. Different scoring process of dimensions to allow a straightforward implementation for the majority of information produced by sensors.

2. Different formal framework, numerical in $[0,1]$ instead of the multivalued logic.

3. Different aggregation process, which is more adaptable in different scenarios.

3.2.2. Graphical representation of the ReCLiC model

The proposed ReCLiC model can be graphically represented as in Figure 3.2 where the central part represents the ReCLiC processing unit and the outer part relates to the inputs and outputs.

It starts from the input on the left which is a log entry l characterised by four components: source ID, date and time, topic and message. The arrows coming from each component indicate that a given dimension relies on this component.

The top part of the figure shows the used meta-information: *TG* represents the state transition graph and *SN* the sensor network. As in the previous case, the arrows point to the dimensions that require this meta-information for proper evaluation.

The bottom part, *H*, represents the history of all previous log entries: after scoring trust for a message, both the log entry and its corresponding trust value are stored. Later this message and its trust value can be used in the evaluation of dimensions pointed by appropriate arrows.

Finally, the central part shows the model itself, together with the respective parameters of its components that will be detailed in the following sections. The box denoted *R* represents reliability, it only depends on the log entry source, its date and on history, as it provides an evaluation of the source depending on its previous assertion. It also depends on a parameter *w*, a temporal time window determining the history horizon to be taken into account.

The second box denoted *Co* represents competence, it depends on the source, its date, the history and, in addition, the topic. It is still an evaluation of the source but with respect to the topic of the message. This aspect is the main difference between competence and previous reliability. The *w* parameter describes, as in the previous case, a temporal time window determining the history horizon to be taken into account.

The third box denoted *Lkl* represents likelihood. Since this dimension evaluates the content of information, it considers the source, the history, the state transition graph (TG), the topic and the message content. Both history and the considered log entry are crucial as they allow for checking the temporal confirmation of the considered information based on the transition graph.

The last box denoted *Cr* represents credibility. It depends on the source, the topic, the message, the history and the sensor network (SN). The history contains messages from all sources thus it is the appropriate place for credibility to search among previous information to seek confirmation or invalidation. It uses the sensor network to filter relevant sources and their topic that can be used in that regard.

All scores from each dimension are then combined in the *aggregation* step which leads to the final trust value. This trust score is then attached to the original piece of information represented by the log entry *l* and saved in the history *H* for later potential usage by either dimension.

3.3. Reliability: initial source evaluation

Reliability is an *a priori* assessment of the source meant to verify its ability to provide any meaningful output on any topic. In this section, various approaches are first discussed to compute it, organised as constant vs. dynamic and meta-data-based vs. history-based. Then a formal definition is proposed, based on the number of error messages that are recently produced as an indicator of whether the sensor is working properly or not. Given a message *l* to be scored, reliability thus only depends on *l.s* and *l.t* to have a dynamic measure evolving across time and *H*, the stored history.

3.3.1. Discussion

A basic approach to measure reliability could make it constant, e.g. only based on the sensor type (or brand). It may be based on some sensor ranking process: for instance, the specification of a camera defines the resolution of its optics which can be better for one than for

another. Thus, one can argue that the first camera *a priori* gives better images without seeing any. The difference between the reliabilities of various devices can thus be based on product specifications. In that case, the highest reliability can be acquired by fulfilling the highest standards. It is important to highlight that it is only relevant to compare the reliabilities of devices that produce a similar type of information.

A way to enrich this basic definition is to take into account time and to define dynamic reliability, for instance considering that this initial reliability value decreases when the sensor becomes older. This approach requires disposing of knowledge about the obsolescence speed of the sensors, which might be difficult to get. This type of information usually comes from the manufacturer in the technical specification. The information provided is the recommended working time for the device which suggests that, if exceeded, the device starts to deteriorate faster, resulting in decreased reliability. Two approaches are possible to include this correlation, either decreasing reliability slowly up to the predefined deterioration point and then speeding the process or assuming high reliability up to the predefined deterioration point and starting reliability decrease afterwards.

Note that it is possible to further enrich this dynamic definition of reliability by taking into account maintenance operations if their dates and types are known. Still, their interpretation can be debatable: they can be seen either as increasing reliability by slowing down the sensor ageing, or they can be considered as the sign of the sensor decreasing reliability due to the need for repairs. The proper identification of the maintenance impact needs to be first experimentally obtained as it may vary with the types of devices. Altogether this information is not easy to obtain.

All these approaches rely on the availability of very rich meta-information about the sensors, such as: type, brand, age, obsolescence speed and dates of maintenance operations. Another source of information that can be exploited to define sensor reliability is the history of its previous outputs, that normally is available in the log file. Indeed, reliability can be related to the question of whether the device is working properly or not, which can be derived from its downtime or its error messages. The next section describes in more detail a reliability definition based on error messages as a base for the proposed scoring method.

3.3.2. Proposed definition

When focusing on sensors, we propose to adapt the general reliability definition in a way that is coherent for most sensor cases and types of devices. The key aspect of reliability is that its evaluation does not depend on the information per se but focuses only on the question: ”can the source provide any meaningful information at the current moment?”. To answer this question we should again highlight that the source, in this case, is a device. The main problem with a device is that it may experience technical faults leading to issues with its output. The high proportion of errors among recent messages on any topic can lead to some doubts about

whether the sensor can currently produce a trustworthy message. In this form, reliability does not consider the message that is being evaluated, it describes the sensor that produced that message either as working properly or not and it serves the purpose of an initial regulator. We propose a dynamical measure which automatically updates the current reliability of the sensor, depending on what happened in its recent history.

Formally, we propose to define reliability as follows:

$$\begin{array}{rcl} r & : \mathscr{S} \times \mathscr{T} & \longrightarrow & [0,1] \\ & (s,t) & \longmapsto & 1 - \dfrac{|error\,(recent_w(s,\mathscr{Z},t))|}{|recent_w(s,\mathscr{Z},t)|} \end{array} \tag{3.1}$$

where $recent_w : \mathscr{S} \times \mathscr{Z} \times \mathscr{T} \to \mathscr{P}(\mathscr{L}_s)$ is the function such that $recent_w(s,z,t)$ returns the set of log entries produced by sensor s for topic z in the considered time window w starting with the time of the message t. As reliability does not depend on the topic, its definition relies on $recent_w(s,\mathscr{Z},t) = \bigcup_{z\in\mathscr{Z}} recent_w(s,z,t)$: this union collects all entries produced by sensor s, whatever their topic is in the considered time window. $error : \mathscr{P}(\mathscr{L}) \to \mathscr{P}(\mathscr{L})$ is the function which extracts the error entries from a set of messages.

The definition of the considered time window, which determines the notion of "recent history" can take several forms: it can be directly defined as a number of entries, indicating the number of previous messages one may want to take into account; it can also be a time interval, from which the log entries to be considered must be retrieved.

The former gives a set with constant number of entries i.e. the denominator $|recent_w(s,\mathscr{Z},t)| = w$, however the time difference between the evaluated entry and the entries in the set can vary from minutes to hours or even days which can falsely impact the scoring due to outdated data. The second approach considers entries from an a priori defined interval. However, there is a risk that the set of entries will be empty which means that the quotient is ill-defined as its denominator is 0, as well as its numerator. In such a case, we propose to set reliability to 1, as there is no indication that the source is not reliable in this case. This can be considered as an optimistic choice, other possibilities can, for instance, consist in setting the reliability to an intermediate level, as 0.5.

By defining reliability this way we do not consider the current message, its source is only characterised as working properly or not. This dimension tests the cases where the device might not work properly, thus potentially giving not relevant output.

Different values of time window are analysed in Section 5.4.4, p. 104.

3.4. Competence: a refined source evaluation

We define competence as relying on both the source and the topic of the considered piece of information, which is coherent with the definition that can be found in the literature (see

subsection 2.1.4, p. 27). As for reliability, competence is a source metric meant to assess the quality of the information provided by this source on this topic. It is similar to reliability but considers that a source can provide information on more than one topic and that, depending on the topic, the information can have different qualities. Given a message l to be scored, competence thus depends on its source, the history and the topic of this message. However, like reliability, it does not consider the information content. If the sensor provides information only on a single topic, competence is the same as reliability. This section first discusses various possible approaches when considering competence for sensors and then formalises our proposed definition in ReCLiC.

3.4.1. Discussion

We view two major approaches when defining competence. In the first case, a competence measure reflects the adequacy between the sensor's operating conditions and its predefined viable working conditions. Such a situation can be illustrated by a kitchen scale being used for measuring a person's weight or in reverse, using a person scale to weigh ingredients for a meal: no matter how reliable the scales are, when they are used outside their scope, they can be considered as incompetent, providing wrong results. Another example is an object recognising system which can only work during the day but not the night.

This type of competence is difficult to assess, it is based on meta-information which is highly dependent both on the type of a device and the environment it is used in. Moreover, considering competence in this form might be irrelevant as most often the type of a device is specifically chosen to fit requirements, thus a situation where a sensor is outside of its working scope rarely happens.

In the second case, a competence measure reflects the different accuracy level between various categories of the message content. Accuracy represents the difference between the information reported by a sensor and the reality. The smaller the difference, the more accurate the sensor is. As an example, let us again consider an object recognition system, instead of analysing its operating conditions, we analyse its ability to recognise different objects. For instance, it can provide more accurate information about planes than helicopters, differencing accuracy between these two cases.

These different cases are called topics, they provide a global indication about a category of the message content (see subsection 2.2.1, p.29). One of the possibilities to score competence is to evaluate how accurate the past results of this sensor for each topic were. This approach can be very effective in some applications, for instance, in the case presented by Mercier et al. (2008) where the system has to recognise one of three objects. The sensor recognises these objects with different accuracies for each, indeed having different competences for the different objects. This approach strongly relies on ground truth to learn a priori how the evaluated sen-

sor manages in different situations which do not correspond to the desired properties of our model, as discussed in Section 3.1.

Another possibility is based on recognising situations where a sensor is especially prone to providing false output. Similarly to reliability, problems can be observed in cases where a device produces error messages. Now, if a sensor provides information on multiple topics, it is possible that only parts of its systems are corrupted, resulting in issues only for a specific topic. For instance, in the case of an object recognition system, parts of this system can be damaged resulting in problems with recognising planes but not helicopters. The next section describes in more detail a competence definition based on error messages as a basis for the proposed scoring method.

3.4.2. Proposed definition

The definition of competence we propose is similar to reliability but more specific, insofar as it performs some topic-filtering: it considers that some of the messages from a given device can be less trusted than others. To apply this idea, the source is assessed only considering the topic of the message it currently provides. In this definition, the content of the information is not considered, making competence a source evaluation.

Whereas reliability considers all recent error messages, competence distinguishes them depending on their associated topics: if only one of them is involved, it means that the source may be able to provide trusted information on other topics. It also means that a high proportion of errors among recent messages for the considered topic can raise some doubts about whether the sensor can currently produce a high-quality message on that topic.

As for reliability, competence is based on a dynamical measure that computes the current competence of the source based on what happened in the recent history; formally it is defined as:

$$\begin{array}{rcl} co : \mathscr{S} \times \mathscr{Z} \times \mathscr{T} & \longrightarrow & [0,1] \\ (s,z,t) & \longmapsto & 1 - \dfrac{|error(recent_w(s,z,t))|}{|recent_w(s,z,t)|} \end{array} \tag{3.2}$$

using the same definitions as for reliability: $recent_w : \mathscr{L} \times \mathscr{Z} \times \mathscr{T} \to \mathscr{P}(\mathscr{L}_s)$ provides the set of log entries produced by sensor s in topic z within the considered time window w starting with the time of the message t, $error : \mathscr{P}(\mathscr{L}) \to \mathscr{P}(\mathscr{L})$ extracts the error entries from the set of recent messages.

The definition of the time window can have various forms, as discussed for reliability in Section 3.3, p. 42. Even though its parameters for competence do not need to be the same as for reliability, we propose to give them the same value. Indeed, we propose to base the definition of competence on a subset of the entries considered for reliability, applying a topic-based filter to these entries, so as to make reliability and competence consistent one with another. Otherwise,

it may be the case that competence depends on entries that are not taken into account for reliability, which may lead to contradictory results.

This choice implies that the denominators of the quotients involved in reliability and competence definitions satisfy the inequality $|recent_w(s,z,t)| \leq |recent_w(s,\mathcal{Z},t)|$. Even if the temporal definition is defined as a number of entries (and not a time interval), it can be the case that the denominator $|recent_w(s,z,t)|$ does not equal w.

It must be underlined that the numerators of the quotients obviously satisfy the same inequality $|error(recent_w(s,z,t))| \leq |error(recent_w(s,\mathcal{Z},t))|$.

Finally, in the case where the denominator is zero, and thus the numerator as well, we propose to set competence to 1, as we proposed for reliability.

3.5. Likelihood: temporal confirmation

In the framework proposed by Revault d'Allonnes and Lesot (2014), the third dimension, named plausibility, aims to compare what is known by the rater with the evaluated piece of information, taking into account his/her background knowledge.

The transposition of this notion of background knowledge to the case of information provided by sensors is not straightforward. We propose to replace the plausibility dimension with a new one, called *likelihood*, that similarly takes into account the information content but combines it with a different notion of context. This proposition, discussed below, makes it possible to achieve the desired aim of making the scoring model independent of ground truth and meta-information.

As previous sections, this section first discusses several possibilities with various degrees of expressiveness for this dimension then it describes the proposed definition.

3.5.1. Discussion

Likelihood is a new dimension we propose to improve assessing information produced by sensors, that has not been introduced before to the best of our knowledge. Likelihood is similar to plausibility, defined in Revault d'Allonnes and Lesot (2014) (see Section 2.1.4, p. 27), as they both take into account the information content and compare it to additional available knowledge. In plausibility, the additional knowledge comes from the rater, it is defined as his/her background knowledge. As such plausibility gives the possibility to personalise the score and adds a subjective compound to the information score. For likelihood, we propose to define this background knowledge as the temporal background, i.e. the history of the considered information to be scored: likelihood is viewed as the compatibility of the information with what happened before and can thus also be interpreted as a form of temporal confirmation. Temporal confirmation means that a piece of information is assessed depending on its date: in general

likelihood answers the question: " is the piece of information consistent with respect to the temporal context?".

It can be observed that the temporal background may refer to external knowledge as ground truth. For instance, detecting a train approach can be correlated with its timetable: if a train is not supposed to pass for another couple hours, then the likelihood of this information might be questioned. Similarly, a thermometer output can be combined with the previous weather forecast: if the two values are significantly different then the likelihood of this information might also be questioned. Both examples rely on ground truth information, therefore they are not compatible with the desired characteristics of the proposed model.

Another interpretation, independent of ground truth, considers the date associated to the considered piece of information not in an absolute meaning, as in the above approach, but in a relative meaning, as a comparison with the previous messages: the considered piece of information is viewed as one element in the sequence of produced messages, about which the state transition graph (see Section 3.1.3.1, p. 37) provides useful information. The next section describes in more detail how this approach can be used to define the likelihood dimension in the case of information produced by sensors.

3.5.2. Proposed definition

As discussed in paragraph 3.1.3.1, p. 37, the state transition graph provides useful information about admissible sequences of messages.

It can be used to understand how the device works and highlight any irregularities whenever they occur. This represents a temporal dependency which verifies whether the current piece of information can occur after the previous one.

We, therefore, propose the following definition of likelihood that depends on the message, its source and topic, where the source and topic are only used to extract its recent history (previous message). To score likelihood, compatibility of this previous message with the current one is obtained from the state transition graph, the dimension is defined as:

$$
\begin{aligned}
lkl \quad &: \mathscr{S} \times \mathscr{Z} \times \mathscr{M} \quad \longrightarrow \quad [0,1] \\
&(s,z,m) \quad \longmapsto \quad \begin{cases} trust(prv(s,z)), & \text{if } prv(s,z) \text{is compatible with } m \\ 1 - trust(prv(s,z)), & \text{otherwise} \end{cases}
\end{aligned} \tag{3.3}
$$

where s is the sensor producing the current piece of information m on topic z, $prv(s,z)$ returns the previous message from sensor s on topic z and $trust(m)$ returns the trust value of message m. Message m_1 is compatible with m_2 only if there is a connection in the state transition graph from m_1 to m_2.

When the information flow is distorted, the likelihood value is all the lower as the trust

score of the previous message is high. This approach does not assign a low likelihood value to the situation where the lack of compatibility results from a piece of information with low trust that should be taken with caution as it may be false.

When the information flow is validated by the state transition graph, a message m can be considered likely with a degree equal to the trust value of the previous message from the same source about the same topic, $prv(s,z)$. This means that if the previous message was not trusted, the current one is not considered especially likely, showing that the validation provided by the state transition graph should be taken with caution.

Providing the correct state transition graph is thus key for scoring likelihood. Lack of transitions that should be present in the graph leads to false alarms. Including a transition which should not be present in the graph results in not recognising specific problems when the associated trust model is used.

It is possible to enrich this process by considering a longer history than a single previous message, i.e. performing compatibility checks for longer sequences. This, however, might unnecessarily make the model more complex. We can notice that the proposed definition is recursive, the trust model depends on the trust model used for scoring previous messages, thus, the trust of the previous message already includes the trust scores of the previous ones.

3.6. Credibility: spatial confirmation

Credibility assesses a piece of information by considering how it aligns with information reported by other sources. It is an essential step, present in the vast majority of scoring models including ReCLiC. Credibility is related to likelihood insofar as both depend on content, rather than the source. More importantly, both look for confirmation: likelihood focuses on a temporal confirmation, provided by previous messages of the same sensor, whereas credibility is based on the spatial confirmation, though the information provided before by other sensors in relation to their geographic position. The approach of how other information can be used to compute credibility depends on the situation and the available knowledge of the system. That makes it difficult to propose a general method that works in many different cases and approaches are usually limited to specific cases (see paragraph 2.2.2.3, p. 31).

In this section, a new approach is proposed to score credibility for information produced by sensors. It is based on the expected correlations between devices in the sensor network where information from one sensor can influence information from the other. In particular, spatial confirmation is considered, based on the distance between devices.

An in-depth description of the approach to score credibility is discussed in the following sections. First, a general model of credibility is defined and discussed in subsection 3.6.1. Then each component is discussed in more detail: the definition of candidate sources in subsec-

tion 3.6.2 and that of confirming or invalidating messages in subsection 3.6.3. Finally, the aggregation process of the extracted messages is discussed in subsection 3.6.4.

3.6.1. General definition

Credibility takes into account the source, the topic and the message. As in the case of likelihood, credibility mainly depends on the provided message, its source and topic are necessary to search for information from other sources that can be used to score it.

The crucial element of our approach is the definition of two functions: *conf* and *inv* that will decompose the set of information provided by correlated sources into confirming or invalidating messages in three steps, described in more details in the following sections.

First, expected correlations have to be defined between all available sensors. Such correlations show that a message from one sensor should be related to a message from the other one. This potential correlation between sensors is derived from a *network of sensors* (see paragraph 3.1.3.2, p. 39), which can be defined based on the geographical location of the considered devices. In this step, the set of candidate sources is extracted.

In the second step, messages from candidate sources can have one of three roles: they can serve as confirmation, invalidation or be neutral. Two functions $conf(s,z,m)$ and $inv(s,z,m)$ are designed to gather messages from the correlated sources that are relevant to scoring credibility of the evaluated message m. The former extracts only the messages that confirm m, the latter finds messages that invalidate m.

In the last step, aggregations are defined. First, the messages representing confirmation and invalidation are combined separately, providing two values. Then, these two values are aggregated to into the final credibility score. Different approaches for aggregation are presented and discussed in the following. The general formula is:

$$\begin{aligned} cr \quad &: \mathscr{S} \times \mathscr{Z} \times \mathscr{M} \quad \longrightarrow \quad [0,1] \\ &(s,z,m) \quad \longmapsto \quad agg_1(agg_2(conf(s,z,m)), agg_3(inv(s,z,m))) \end{aligned} \tag{3.4}$$

where m is the evaluated message produced by sensor s on the topic z. The function $conf(s,z,m)$ returns a set of messages that confirm m; the function $inv(s,z,m)$ returns a set of messages that invalidate m. agg_1, agg_2 and agg_3 are three aggregation operators that return a single value by analysing a set of messages or values that represent them (e.g. the trust score).

3.6.2. Sets of candidate sources

To find confirming or contradicting messages from other sensors, the first step is to extract the subset of sources that can be correlated with the source that produced the considered message. This correlation means that an event that impacts the first source may also impact the correlated one which results in messages from both sources. For this purpose, the *network of*

sensors is used (see paragraph 3.1.3.2, p. 39) where a node represents a source and connections between nodes represent expected correlations. Using this graph, it is straightforward to extract a subset of correlated sources, as the neighbours of the considered one.

To build the *network of sensors*, we propose to consider their geographical position with respect one to another: in the proposed approach based on the spatial confirmation, a correlation is expected when sensors are close to each other. As sensors transform a part of reality into a digital value, they are usually close to the event or the situation they report on. Because of that, in many cases, sources close to each other are expected to manifest correlative behaviour when analysing the information they provide. For instance, two thermometers should provide similar readings when they are located close one to another. When the distance increases, a correlation also exists but it might be less important, i.e. the readings can have larger differences.

The notion of topics allows to improve recognising correlations between sources. In particular, the situation where two sensors are close but they are not expected to be correlated, because they apply to different topics. For instance, when considering the thermometer case, two devices might be close to each other but one provides information about temperature inside a building and the other outside a building. These are two different topics thus the two devices are not expected to provide correlated information.

3.6.3. Sets of confirming and invalidating messages

After choosing the correlated sources, the next step is to propose the type of messages that can be used as confirmation or invalidation in a specific situation. For numerical messages, this step is often straightforward as it is possible to compute and interpret the difference between the two values. For event-type messages, this step is challenging since there is no direct connection between different types of messages from different types of sensors. This requires to specify the exact conditions when a message from an external source can confirm or invalidate the evaluated information.

To specify which messages can be considered as confirmation or invalidation for the evaluated piece of information, a tuple (s_t, m, w) is defined. It describes every available piece of information from correlated sources. In that tuple, s_t describes the type of the source, m is the message and w is a time window. Tuples are defined a priori to describe the necessary attributes of the message produced by the correlated source to be considered in credibility computation. Since at this step the messages are only provided by correlated sources, the sensor ID and the considered topic are not included in this tuple.

The type of the sensor, s_t, is necessary as different devices produce different messages. This is especially important in numerical messages as, for instance, the value 10 means something different when produced by a thermometer or a barometer. The message m defines what message is necessary to be used as confirmation or invalidation and the time window w sets a limit

to the time difference between the evaluated information and the message considered for its confirmation.

Two sets of tuples are considered, one that represents the messages that confirm the evaluated piece of information and a second one that gathers the messages that invalidate the evaluated piece of information. The messages that do not fit in tuples of neither confirmation nor invalidation are considered neutral and they are omitted in the credibility scoring process. This neutral case also includes all error-type messages: their role is different, they are involved in the definition of reliability and competence, but not considered as confirming, nor invalidating.

To illustrate this idea, let us again consider a thermometer. For each temperature, a set of tuples, used to analyse the messages provided by correlated sensors, is created. Let us take temperature 10° as an example of the evaluated message. Since in this example we only use other thermometers to seek confirmation, this is our only type of the source s_t = *'thermometer'*. The considered messages should be close to the 10°, let us consider only the closest values for simplicity, e.g. 9°, 10° and 11°. The last argument defines the time after a message from an external source can no longer be considered useful, in this case, it can be set e.g. for 10 seconds $w = 10$. Finally we define three tuples (s_t, m, w) to confirm the evaluated message of 10°: $(thermometer, 9°, 10s)$, $(thermometer, 10°, 10s)$ and $(thermometer, 11°, 10s)$. This means that if a message from a correlated sensor, produced less than 10 seconds ago, has value 9,10 or 11 and the type of the correlated source is *thermometer*, then this message is treated as confirmation.

3.6.4. Aggregation

All relevant messages, both confirming and invalidating the evaluated information, need to be aggregated to be used later in scoring credibility. Choosing a single suitable function for this aggregation is difficult as the type and quantity of available messages may vary, depending on the considered domain and devices that produce them. Two major problems can be highlighted for this step. First, we have to interpret the influence of each message depending on the available information, then we need to choose a strategy to combine them. In this section, several propositions are presented about how to define agg_2 and agg_3 showed in Equation (3.4), p. 50.

3.6.4.1. Complexity levels

Various approaches can be considered depending on the available information included in a message. Several complexity levels are discussed below. We denote $\mathcal{M}$ the set of messages to be aggregated, either for confirmation or invalidation of a message m_0. One can for instance use:

– the number of messages $|\mathcal{M}|$

– the associated quality scores (e.g. trust) $\{trust(m), m \in \mathcal{M}\}$

– additional reinforcements $\{(trust(m), d°(m, m_0)), m \in \mathscr{M}\}$, where $d°$, for instance, denotes the degree of confirmation or invalidation

In the most basic approach, the absolute cardinality $|\mathscr{M}|$ can be compared to the total number of candidate messages before their classification as confirming, invalidating or neutral. The outcome, in the $[0,1]$ interval, describes the influence of either side, confirmation or invalidation. Since the neutral set can be very large, it is possible to limit the relation from all messages to the sum of confirming and invalidating sets. In this case, a basic comparison can be proposed to score credibility.

At the next level, additional information can be taken into account: each message in $|\mathscr{M}|$ can be associated with its own trust value. When considering these messages for aggregating confirmation or invalidation, it is then possible to use this additional trust score to improve the credibility computation. by differentiating between messages with low trust and decrease their influence in the final score. In this case aggregation consist in combining a set of trust values, i.e. it is a function $[0,1]^{|\mathscr{M}|} \rightarrow [0,1]$ of qualified messages.

These propositions can be further extended to distinguish between partly confirmed (invalidated) message from fully confirmed (invalidated) one. This approach is based on additional knowledge indicating the extent to which a given piece of information confirms or invalidates the evaluated piece of information. This degree can be correlated with the network of sensors where weights between nodes can suggest the level of correlation between sensors. The higher the correlation, the higher is the degree of confirmation or invalidation between messages from these sensors (see e.g. Lesot et al. (2011)). However, in the following, this approach will not be discussed further because the exact correlation between sources as well as its influence on the messages produced by them is a difficult topic and requires further research. A major difficulty is its high dependence on the situation at hand, which makes it unfit for the generic method desired in this thesis.

3.6.4.2. Choice of aggregation operator

The next step is to choose the aggregation operators: after a short reminder about the main existing families, we discuss the first two levels of complexity introduced in the previous section in turn.

Short reminder about some aggregation operators There exists a very abundant literature about aggregation operators, that offers a wide range of behaviours (see Detyniecki (2001) for instance): an aggregation operator is generally defined as a function taking as input a set of numerical values, usually in $[0,1]$ and outputting a single value, also in $[0,1]$. One can, for instance, distinguish conjunctive, disjunctive and compromise operators: the conjunctive ones, also called t-norms, output a value that is lower than the minimal input value. As a consequence their output is large if all the aggregated values are high, hence their name of con-

junctive operators. When applied to two input values x and y, they for instance include the functions $\min(x, y)$ or $x \cdot y$.

On the contrary, disjunctive operators, also called t-conorms, output a value that is higher than the maximal input value: their output is large if at least one from the aggregated values is high, hence their name of disjunctive operators. When applied to two input values x and y, they for instance include the functions $\max(x, y)$ or $x + y - x \cdot y$.

In between, the family of compromise operators allow for trade-off behaviours, where some low values among the ones to be aggregated can be compensated for by high values. This category can be illustrated by the average operator, as well as its weighted variant or the order-weighted one. The latter, called OWA (see Detyniecki (2001)), attaches weights that depend on the rank of the values when they are sorted and allows to define a continuum between the minimum operator (where a single non-zero weight is attached to the value with the largest rank when sorted in decreasing order) and the maximum operator (where a single non-zero weight is attached to the value with rank 1).

Basic scenario If only $|\mathcal{M}|$ is available, the computation is limited to the number of messages that confirm or invalidate the evaluated piece of information. In this case, we can propose that at least x messages are necessary to fully confirm the information. To use OWA aggregation in this approach, trust for each message is constant and set to 1. If there is less than x required messages, absent messages are represented as 0 for OWA computation.

Using the OWA operator allows to offer various behaviours in the case when there are not enough messages to fully confirm the information. For instance, if the required number of messages is four, the highest confirmation is obtained only if four or more messages are present. If there are less than four messages, the confirmation is decreased. The level of this decrease can depend on the desired characteristics. For instance, if at least one message is necessary to obtain high confirmation, one can set high weight for the first message and lower ones for the other ones, e.g. $W = [0.8, 0.1, 0.05, 0.05]$. Then if only two messages are available i.e. if $|\mathcal{M}| = 2$, the confirmation level is 0.9. On the opposite, one can highlight that even one missing message results in a significant decrease in confirmation which can be done by setting the last message with the biggest weight: $W = [0, 0, 0.2, 0.8]$. Even having three messages, the confirmation is only 0.2. This approach can be used both for confirming and invalidating messages, respectively: agg_2 and agg_3

Aggregating trust values In the refined case that exploits the available trust scores, the OWA aggregation operator allows to sort messages to consider first the ones with high trust scores and increase their influence in the final score. This behaviour is expected to base the score on the most trusted messages as opposed to the ones with low trust value which might falsely contribute to either confirm or invalidate the message.

As in the basic scenario, we can propose an x threshold which describes the number of messages required to consider the information as fully confirmed. This approach might be useful when the total number of possible confirming messages is unknown or if there is only need for a few of them to confirm the evaluated message. The corresponding OWA operator is associated to the weights $W = [w_1, w_2, ..., w_x, 0, 0]$ applied to the trust values ordered from the highest to the lowest, discarding the messages after the x threshold. In the basic form, the trust average of x messages can be extracted, setting weights as: $W = [\frac{1}{x}, \frac{1}{x}, ..., \frac{1}{x}, 0, 0]$. Combining OWA weights with trust values of the considered messages can result in different scenarios. For instance, three messages with low trust can provide lower confirmation than one high trust message. This approach can also be mixed with the propositions in the basic scenario to highlight the importance of at least one message or expect all of them to have high confirmed information.

3.6.5. The final credibility aggregation

After obtaining the confirmation and invalidation scores, the last step aims at combining them into the final credibility value, formally performed by function agg_1 in Equation (3.4), p. 50.

There are multiple approaches to combine these two values. In this section, we discuss several such approaches depending on the characteristics of the task at hand and the user objectives.

3.6.5.1. Discussion

The final credibility aggregation is based on two values returned by agg_2 and agg_3, both in $[0,1]$ range. Thus $agg_1 : [0,1] \times [0,1] \rightarrow [0,1]$ represents the influence of confirming and invalidating messages to the final credibility score.

The way they are aggregating strongly depends on the user's needs and the availability of confirming and invalidating messages. We consider four possible types of behaviour requirements.

- default credibility has a medium value that can either be increased by confirmation or decreased by invalidation,
- default credibility is high and can only be lowered by invalidation,
- default credibility is low and can be increased by confirmation,
- default credibility has a neutral value that can be increased by confirmation and decreased by invalidation.

The difference between *medium* and *neutral* credibility is that the former has a starting value in the middle between the lowest possible credibility and the highest one. The neutral credibility can be applied when confirmation and invalidation sets are empty. In such a case, credibility

should be viewed as not defined until any confirming or invalidating messages are present. The neutral value is presented by Revault d'Allonnes and Lesot (2014) where their extended multi-valued logic is able to incorporate an undefined value of a dimension. Thus the fourth case cannot be represented by a simple numerical scale of $[0,1]$ range.

3.6.5.2. Proposed aggregation

The final aggregation, denoted agg_1, combines the two values obtained in the previous sections: the level of confirmation, denoted $c = agg_2(conf(s,z,m))$, and the level of invalidation, denoted $i = agg_3(inv(s,z,m))$. Both c and i are within the $[0,1]$ range, as well as the results of the final aggregation.

If only one of the two values is available and the other one cannot be obtained, there is no need for additional aggregation. In this case, $cr = c$ if only confirmation is available and $cr = (1-i)$ if only invalidation is available. Indeed, it is necessary that an increase in invalidation results in a decrease in the final credibility score. This behaviour is actually required whether the confirmation is available or not: here as well as in the following, the two considered quantities are c and $1-i$ (and not i).

If both c and i are available, the final aggregation depends on the user's needs. Three different aggregation types are considered, as briefly introduced in paragraph 3.6.4.2: compromise, conjunctive and disjunctive. They are represented here by three basic formulas, average for compromise aggregation, probabilistic operators for conjunctive and disjunctive aggregation, formally:

- compromise behaviour: average $agg(x_1,x_2) = \frac{x_1+x_2}{2}$
- conjunctive behaviour: probabilistic t-norm $agg(x_1,x_2) = x_1 \cdot x_2$
- disjunctive behaviour: probabilistic t-conorm $agg(x_1,x_2) = x_1 + x_2 - x_1 \cdot x_2$

The compromise approach is proposed by default as it equally considers confirmation and invalidation. It starts from the neutral point $\frac{1}{2}$, if the evaluated piece of information is neither confirmed or invalidated. Then, confirmation can increase this value and invalidation decrease it. Formally, it is defined as:

$$agg_1(c,i) = \frac{c+(1-i)}{2} \tag{3.5}$$

In this example, the credibility score depends on whether the confirmation or the invalidation score is stronger. When both $c = 0$ and $i = 0$, the neutral point is met. When the confirmation score increases, credibility increases as well; when the invalidation score increases, credibility decreases.

In the conjunctive approach, the evaluated information is considered credible only if it is both confirmed and not invalidated. This is the most sensitive approach since not fulfilling

both conditions results in a fast decrease in credibility. Formally, it is defined as:

$$agg_1(c,i) = c \cdot (1-i) \tag{3.6}$$

An absence of confirmation ($c = 0$) results in credibility scored as 0, which is the same result when the information is fully invalidated. This approach is used when no contradictory messages are allowed. The credibility score is largely decreased if there are any indications that the evaluated message might not be credible, either because it was not fully confirmed or some invalidation was encountered.

In the disjunctive approach, the evaluated information is considered as credible if it is either confirmed or not invalidated. This means that credibility is low only when the information is invalidated and not confirmed. This approach discards any invalidating messages as long as the confirmation is high. Formally, it is defined as:

$$agg_1(c,i) = c + (1-i) - c \cdot (1-i) \tag{3.7}$$

In this approach, even strong invalidation is disregarded as long as there is strong confirmation as well, resulting in high credibility. It can be implemented in situations where confirmation is more important and invalidation is considered only when no confirmation is available.

3.7. Trust

The final step to assess the trust to be put in a message provided by a sensor in the ReCLiC model consists of aggregating the four previous dimensions: reliability, competence, likelihood and credibility. Again, this step allows for a large variety of approaches, which makes the ReCLiC model highly flexible to be adapted to the different needs of the user.

For this study, three major operators are considered, similarly as in the credibility case discussed in the previous section: compromise, conjunctive and disjunctive aggregations. In the following, they are considered in different combinations in order to highlight different features: after discussing these combinations, this section describes the recommended one.

3.7.1. Discussion

As mentioned earlier, three main approaches to aggregation can be used to apply different effects when joining pairs of dimensions. However, considering all combinations of dimensions and aggregations results in too many configurations and can introduce unnecessary redundancy where one combination is very similar to the other. In order to reduce the number of variables, some additional requirements are discussed.

Proposed restrictions First, it is important to highlight the similar properties of reliability and competence: they both describe the source, to evaluate a priori its ability to produce relevant pieces of information. Due to this characteristic, it is recommended to consider them together. Furthermore, to highlight any decrease in either of them, a conjunctive behaviour can be considered as relevant: in the following, we thus consider the product of reliability and competence (using the probabilistic t-norm, see paragraph 3.6.4.2, p. 53) as a single dimension.

For the further analysis, a schema is proposed to divide all dimensions into two categories whose importance can be adjusted. This adjustment is based on the weighted average that can balance one side over the other. It is represented as:

$$trust = \alpha \cdot agg(d_1, d_2) + (1 - \alpha) \cdot d_3 \tag{3.8}$$

where d_i are dimensions: reliability combined with competence ($r \cdot co$), likelihood (lkl) or credibility (cr) and α is a constant that regulates the importance of either side. First, the position for each dimension is chosen, then the type of agg is determined for d_1 and d_2.

This is a two-step approach where the first two dimensions are aggregated resulting in a preliminary trust score which is later either increased or decreased by the remaining dimension. The fixed inclusion of compromise at this level ensures that the proposed framework is neither too pessimistic (only conjunctive aggregations) nor optimistic (only disjunctive aggregations).

The aggregation $agg(d_1, d_2)$ can have one of three main forms discussed in the previous paragraph 3.6.4.2, p. 53:

- conjunctive behaviour: probabilistic t-norm $agg(d_1, d_2) = d_1 \cdot d_2$
- disjunctive behaviour: probabilistic t-conorm $agg(d_1, d_2) = d_1 + d_2 - d_1 \cdot d_2$
- compromise behaviour: average $agg(d_1, d_2) = \frac{1}{2} d_1 + \frac{1}{2} d_2$

Overview of 9 trust models When combining the two choice levels, a total number of nine trust models can be defined. Each configuration is denoted v_{ij} where i encodes the aggregation order of the three dimensions and j the type of the aggregation operator:

- $i = 1$: $d_1 = lkl$, $d_2 = cr$, $d_3 = r \cdot co$
- $i = 2$: $d_1 = r \cdot co$, $d_2 = lkl$, $d_3 = cr$
- $i = 3$: $d_1 = r \cdot co$, $d_2 = cr$, $d_3 = lkl$
- $j = 1$: conjunctive behaviour for agg
- $j = 2$: disjunctive behaviour for agg
- $j = 3$: compromise behaviour for agg

	Conjunctive (j = 1)	**Disjunctive (j = 2)**	**Compromise (j = 3)**
i = 1	$\alpha \cdot lkl \cdot cr$ + $(1-\alpha) \cdot (r \cdot co)$	$\alpha \cdot (lkl + cr - lkl \cdot cr)$ + $(1-\alpha) \cdot (r \cdot co)$	$\alpha \cdot (\frac{1}{2} lkl + \frac{1}{2} cr)$ + $(1-\alpha) \cdot (r \cdot co)$
i = 2	$\alpha \cdot (r \cdot co) \cdot lkl$ + $(1-\alpha) \cdot cr$	$\alpha \cdot ((r \cdot co) + lkl - (r \cdot co) \cdot lkl)$ + $(1-\alpha) \cdot cr$	$\alpha \cdot (\frac{1}{2}(r \cdot co) + \frac{1}{2} lkl)$ + $(1-\alpha) \cdot cr$
i = 3	$\alpha \cdot (r \cdot co) \cdot cr$ + $(1-\alpha) \cdot lkl$	$\alpha \cdot ((r \cdot co) + cr - (r \cdot co) \cdot cr)$ + $(1-\alpha) \cdot lkl$	$\alpha \cdot (\frac{1}{2}(r \cdot co) + \frac{1}{2} cr)$ + $(1-\alpha) \cdot lkl$

Table 3.1: ReCLiC final aggregation step: all 9 v_{ij} possibilities.

All 9 combinations are presented in Table 3.1, where rows represent different aggregated dimensions (i) and columns show how they are aggregated (j). The nine versions allow to simulate different behaviours that may be expected by a user. To study and characterise the differences between these variants, an experimental analysis is performed in Chapter 5 where trust dynamics are plotted and their behaviours compared. A default version is discussed in the following section.

3.7.2. The final trust aggregation

Among the multiple versions, the proposed definition used in ReCLiC model consider a conjunctive behaviour of reliability, competence and likelihood with later added credibility that corresponds to configuration v_{21}:

$$trust = \alpha \cdot (r \cdot co) \cdot lkl + (1-\alpha) \cdot cr \tag{3.9}$$

This model separates the three dimensions that solely depend on the considered sensor, from the dimension that also takes into account other, related, sensors. The α value in this configuration allows to move the importance to one of the two sides. When considering low α, the trust score is based more on external validation. If the α value is high, then the more important aspect becomes the internal evaluation.

Furthermore, reliability, competence and likelihood do not consider external sources for their scoring. A conjunctive behaviour represents the intuitively appealing idea that if at least one of the dimensions reports a problem, then high values of the other ones should not mitigate the discovered problem. If a device does not work properly, and thus the reliability or competence value is low, then the likelihood value should not be considered relevant, as the user cannot be certain that the piece of information considered by likelihood is valid. Reciprocally, if the user knows that a piece of information is not likely, then the knowledge that the

source is fully reliable and competent is not that relevant.

Credibility is independent of the internal state of the sensor represented by the other dimensions. Because of that, it serves the purpose of increasing or decreasing the preliminary trust score, depending on whether the message can be confirmed or invalidated.

3.8. Summary

This chapter presented the ReCLiC model for information scoring in the case the considered information is produced by a sensor, with the aims of being sensor-generic, not dependent on ground truth and dependent only on easy-to-access meta-information. Following the principles of state-of-the-art approaches, ReCLiC is a multi-criteria aggregation with various degrees of liberty, for each of the criteria and their aggregation, showing its flexibility and adaptability to the user's needs.

The next chapter studies the implementation of this general model for a specific sensor, called axle counter, used in the railway signalling domain.

4. ReCLiC implementation for the railway signalling domain: the case of Axle Counters

The proposed ReCLiC model for scoring information produced by sensors has been described in the previous chapter from a generic point of view, so as to be applicable to any type of sensor network storing their outputs in a log file.

This chapter describes its implementation to a real case, in the railway signalling domain, provided by the industrial collaboration with Thales in which this thesis took place: information scoring in railway systems is crucial, to ensure safe train passages, improve efficiency in train traffic and decrease possible delays. Furthermore, this domain possesses the characteristics that governed the development of the ReCLiC model: high-level ground truth about train movements is not stored and no ground truth is available at a detailed level to check the individual output of the sensors.

In this chapter, the railway signalling domain is introduced with a discussion on the available types of sensors. One of them, called Axle Counter, is chosen for the ReCLiC implementation to obtain trust values for the produced messages. Each step of the four dimensions scoring is discussed in detail to illustrate a possible usage of ReCLiC model leading to an operational instantiation of ReCLiC that allows its experimental study in Chapter 5. In addition, this chapter offers a formal study of the implemented ReCLiC behaviour, analytically examining the effect of each dimension separately.

4.1. Devices in the railway signalling domain

This section presents the application domain we consider, first giving a global view, then describing the dataset made available by the thesis industrial partner, Thales Polska. It later details the studied sensor, named axle counter, as well as the motivation for this choice.

4.1.1. Overview

Principle Train traffic includes various systems and devices which ensure safe and efficient train travel between stations. Trains travel on tracks which are divided into sections, these sec-

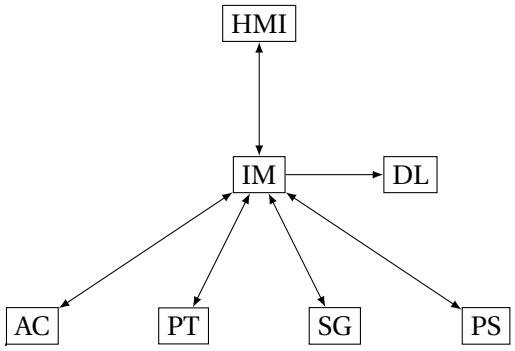


Figure 4.1: An example of the Interlocking Module system

tions include various sensors that allow for better traffic management, especially at a large scale. The most common devices are:

- axle counter (AC): a sensor indicating the presence of a train in a certain part of tracks,
- point machine (PT): a device used to switch tracks at crossroads,
- signal (SG): a device reporting changes in the colour of traffic lights or their failures,
- power supply (PS): a device reporting activities and failures in power transmission,
- data logger (DL): a device responsible for transmitting and storing logs from sensors.

All these devices are connected by a system that processes the messages they produce, executes instructions from a user, displays information and stores it, as described below. This so-called interlocking module (IM) is in the centre of the system, its aim is to provide a connection between the devices and a user.

Any information reported by devices is sent to the interlocking module which stores them using the data logger (DL) module and displays the readings to the human-machine-interface (HMI). A user can react to the reported events by sending commands to IM e.g. setting train routes or fixing problems. The interlocking module interprets these commands and activates the necessary actions by sending instructions to the appropriate devices. Any user's command is also reported in the data logger by IM.

All activities of sensors are stored in a database. The main purpose of this database is to safely store all knowledge about the events that can be observed on tracks, which includes any train passage with as many details as possible. These data are necessary for the situation when an accident happens that involves a train passage: details of the state of each device as well as the operator's commands can then be analysed to find out what caused the event.

Since the lack of this type of information after an accident can result in financial penalties and reputational damage, it is important to ensure its availability and maintain the high quality of the stored information. One of the possibilities is to monitor the information quality in the considered database or in real-time, to find any emerging quality problem and attend them on

the go. This type of monitoring system can be based on the ReCLiC model that can highlight any quality problems in real-time.

Collected data The database we consider in this thesis is called MoTRicS2015. It incorporates log entries produced by devices such as the ones described above, reflecting all activity performed by these sensors. The events stored in this database can be asynchronous: logged after an event happened or a result of a process started by a user.

The MoTRicS2015 database consists of log entries from sensors located in five small Polish train stations with low traffic, covering a period of about one year, between March 1st 2015 and March 17th 2016. The log entries are generated by different sensors located in different parts of tracks as the ones described in the previous section.

The 175GB dataset contains almost 540 million entries. An entry has at most 13 fields, defined for each sensor according to the action it takes, not all of them are easy to understand: for some of them, the purpose is clear e.g. log id, time of the activity or id of the device, for others, the meaning of values can be difficult to understand.

In addition to this database, two types of meta-information are available, as required by the ReCLiC model. The network of sensors can be derived from the plans of stations that are available for the tracks as described in more details in subsection 4.2.4, p. 71. As each station comes along with its own network of sensors, this thesis focuses on one station (name omitted for confidentiality reasons). The state transition graph is is also available, as described in more details in subsection 4.2.3, p. 68.

4.1.2. Axle Counters

As mentioned in the previous section, an axle counter (AC) is a device that informs about a train passing between two points on a track. Its principle is illustrated in Figure 4.2 and described below. Axle counters are event-based, asynchronous sensors with respect to the typology discussed in subsection 2.2.1, p. 29. More precisely, they can be described as sensors that aggregate the data produced by more basic sensors, namely *detection points*, as described below. This section also presents the messages AC can produce.

Principle An axle counter provides information about whether a train entered or left a section. A section refers to a portion of tracks of various lengths. A *detection point* (DP) is installed at each end of the considered section. When wagon axles pass the detection point at the start of the section, a counter is incremented. As the train passes throughout the section, it encounters a similar detection point at the end of this section. The counter compares the counts at the end of the section with that recorded at the beginning. If the two counts are the same, the section is presumed to be clear for a second train.

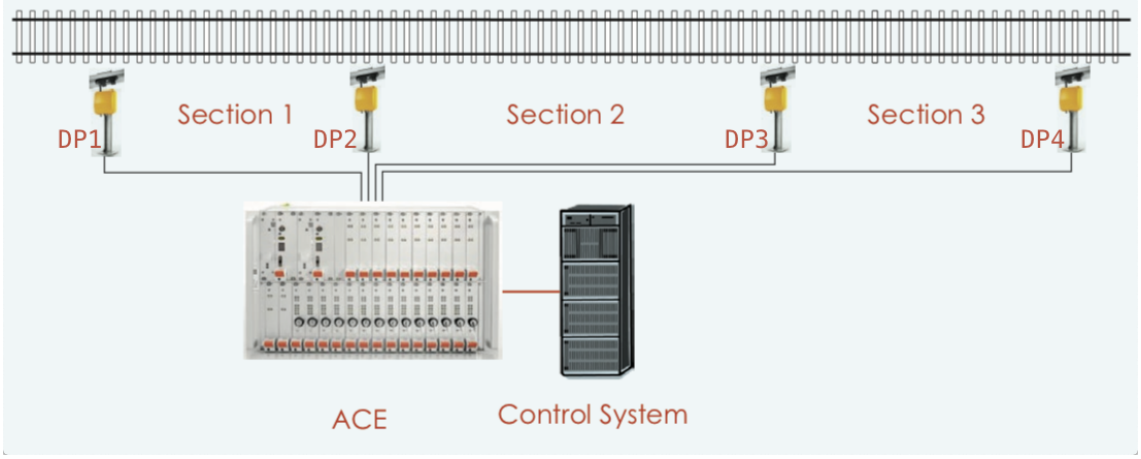


Figure 4.2: Illustration of the axle counter principle which consists in one ACE and multiple DPs. Provided by Thales: https://myproducts-thales.com/brochures-documents/1-brochure-fieldtrac-6315-az-lm/file

The interpretation of counters from detection points is carried out by safety-critical computers called *axle counter evaluators* (ACE) which are usually located in the middle of the sections. The detection points are either connected to the evaluator via a dedicated copper cable or via a telecommunication transmission system. The latter allows the detection points to be located at significant distances from the evaluator which can increase the number of DP that can be connected to one ACE. To summarise, one axle counter has a combination of multiple DPs and one ACE.

Output messages Axle counter can produce five main types of messages, described in turn below:*occupied*, *clear*, *disturbed*, *reset_ rejected* and *reset_ success*. Two detection points are necessary to obtain information about the state of a given section, i.e. whether it is occupied by a train or not. When the first DP increases its counter for the first time, the ACE interprets it as the fact that the train just entered the corresponding section, say *S1*. At this moment, ACE sends the message *occupied* to the control system, including the name of the section that is now occupied by a train as the associated topic.

As the train is passing, more axles trigger the DP and increase its counter. When the train reaches the second DP for the first time, the ACE interprets it as the fact that the train entered the neighbouring section of *S1*, say *S2*. It then sends the *occupied* information to the control system about the topic *S2*. By passing through the second DP, the axles of the train increase its counter. When the value of the second counter becomes the same as the first one, it means that the train left *S1* and ACE sends the information *clear* to the control system about topic *S1*.

We can observe that due to this aggregation step, a section is a concept made from the data provided by at least two DP and their counts interpreted by ACE. Moreover, the output of an AC can be interpreted as a piece of information whereas a DP would only output data, with regards to the discussion at the beginning of Chapter 2, p. 21.

If the second counter has a value larger than the first one, an error occurs and the ACE sends the information *disturbed* to the control system. It then waits for the recovery process launched by a human operator. Note that such a recovery process is not only launched after the *disturbed* message: the human operator can also decide to launch one if he/she observes that the second counter has a value lower than the first one after the entire train has passed. Indeed in such case, the state of the section erroneously remain *occupied* : the role of the operator is to recognise this issue.

The recovery process starts with the message *reset* and can either be rejected, which is logged as *reset_ rejected* or it can be successful, logged as *reset_ success*.

In the chosen station from the MoTRicS2015 database, there are 57 sections handled by 5 axle counters. They provide 1,142,302 log entries within one year time, between March 1st 2015 and March 17th 2016.

4.1.3. Motivation for scoring trust in Axle Counters

The reliable detection of trains is fundamental to the safe operation of railways. This puts a great demand on the availability of train detection systems. In crucial areas, 100% availability is indispensable, for example on lines with a high traffic density, on high-speed lines, on key metro lines, in key railway junctions or tunnels.

However no device is always fully reliable, axle counters are also susceptible to a number of potential problems. For instance, the quality of the electrical signal transmitted by the rail is dependent on the insulation resistance of the ties and the ballast [1]. If insufficient, it causes current leakage. Moreover, the connection between DP and ACE can be compromised and the transmissions may be susceptible to interferences which can distort the message.

Because of these potential problems, it is important to be able to recognise situations when an axle counter, or one of its parts, fails to provide relevant information. It is especially important for axle counters since not knowing the positions of each train can have catastrophic consequences, possibly leading to collisions.

The next section illustrates how the ReCLiC model can be used to assess the information provided by axle counters and how to implement it, to score trust for the crucial messages *occupied* and *clear*. Low trust for any of these messages indicates that there is a high possibility that the reported event did not happen or the information about this event is in some way corrupted.

The information scoring process is limited to the two messages *occupied* and *clear*. Other messages correspond to situations that are necessarily checked by an operator: they are manually processed anyway, independently of a potentially associated trust value. Thus it does not seem necessary to compute such a trust score. On the other hand, low trust scores for the mes-

[1] More information based on Thales products: https://myproducts-thales.com/brochures-documents/1-brochure-fieldtrac-6315-az-lm/file

sages *occupied* and *clear* make it possible to draw the attention of the operator to messages that require checking and that would perhaps not have been checked otherwise.

4.2. Adaptation of the ReCLiC model to the AC case

This section discusses the implementation of the ReCLiC method proposed in Chapter 3, discussing, in turn, each of the four dimensions: reliability, competence, likelihood, credibility as well as their final aggregation. The definition of the required meta-information is also presented in details in the sections describing the dimension they are used for (i.e. likelihood for the state transition graph and credibility for the sensor network).

4.2.1. Reliability

The reliability dimension describes the current state of the source, whether it can be considered trustworthy or not. The approach described in Section 3.3, p. 42 proposes to evaluate it based on recent error messages.

Two aspects are important when implementing the proposed approach: defining the type of messages that correspond to errors and the notion of recent messages, through the definition of the time window w.

Definition of error messages In the case of axle counters, a particular message is produced when a problem occurs, named *disturbed*. After its occurrence, a sequence of user manually launched actions is performed to reset the sensor to enable its further normal work. Even though such a situation is bound to happen, it is undesirable and indicates problems with the sensor: the following messages provided by this sensor should be treated with caution and *disturbed* can be used as an error message.

Definition of the time window As mentioned in Section 3.3, p. 42, the notion of time window can be defined in two ways, either as time range or as a fixed number of messages. The latter case is considered. Indeed, the problem with the former is that information from AC is often produced scarcely, sometimes with no messages for hours. This can often create situations where a single previous message would be considered, if not none. If this message is a disturbed message, it decreases reliability significantly and more generally the assessment of reliability may be considered as not robust enough.

When searching for the *disturbed* messages, the number of considered recent messages is an important parameter: it determines the level of reliability decrease when a new error message is encountered (see details in the formal study of this parameter in subsection 4.3.1, p. 77). A larger value represents the state of the source in longer periods, a smaller one can decrease reliability significantly in a short time. In the case of AC, we propose to set the time window

to 20 previous messages as will be justified experimentally in more details in subsection 5.4.4, p. 104: these experiments show that the value does not decrease reliability significantly with only one or two disturbed messages, which is the desired behaviour: as the occurrence of two error messages should not indicate significant problems with sensors. However, when the number of errors is greater, reliability decreases more drastically and can affect the final trust value significantly.

Final definition When considering the proposed characteristics, the reliability defined in Equation (3.1), p. 44 can be simplified in the AC case to:

$$r(s,t) = 1 - \frac{e_s}{w} \tag{4.1}$$

where e_s is the number of *disturbed* messages produced by sensor s within the last w log entries. When there is no error messages, the reliability value is 1. The range of the e_s value is $[0, w]$.

4.2.2. Competence

Similarly to reliability, competence also evaluates the source. However, in this case, the ability of a sensor to provide correct information on a specific topic is scored (as discussed in Section 3.4, p. 44). The reason for considering topics separately is that only a part of a device can be damaged where the other components are intact and can provide relevant information.

Definition of topics In the case of axle counters, one device, an axle counter evaluator (ACE), is responsible for interpreting the low-level data provided by several detection points (DP). Each pair of detection points is responsible for providing information about a single section where one axle counter can produce information about many railway sections (see details in subsection 4.1.2, p. 63). These sections can be treated as topics in the ReCLiC model since disruption of one DP leads to problems with at most two sections. Because of that, the information from other topics (sections) should not be affected and the competence is only decreased for the one topic that is causing the disruptions.

We propose an approach similar to reliability to compute competence where the lack of competence is considered as the inability of a sensor to provide relevant information for the specific topic. This situation takes place when a sensor starts to produce error information only for this considered topic and the more error information is produced recently, the less competent the sensor is. As in the reliability case, the *disturbed* message is considered as error information.

Final definition The adaptation of Equation (3.2), p. 46 is thus straightforward: considering the same notion of time window as for reliability leads to:

$$co(s,z,t) = 1 - \frac{e_{s,z}}{w_{s,z}} \tag{4.2}$$

where $e_{s,z}$ is the number of *disturbed* messages produced for section z within the $w = 20$ recent log entries from sensor s and $w_{s,z}$ the total number of messages produced for section z within the last w entries of sensor s. In other words, the value $w_{s,z}$ represents the cardinality of the subset of the 20 last messages limited to the ones with section z. The value $e_{s,z}$ provides the number of *disturbed* messages among them. The range of $e_{s,z}$ is $[0, w_{s,z}]$ and the range of $w_{s,z}$ is $[0, w]$.

As discussed in subsection 3.4.2, p. 46, if there is no previous messages within the considered time window about section z, i.e. $w_{s,z} = 0$ (and thus $e_{s,z} = 0$ as well), making the quotient undefined, we propose to set competence to the highest value 1, as there is no indication that the source is incompetent in this situation. This can be considered as a optimist approach, we leave as a direction for future research the definition of other behaviours, through the definition of other default values.

4.2.3. Likelihood

Likelihood focuses on the information itself, in relation to the messages provided in the past by the same sensor on the same topic. The approach presented in Section 3.5, p. 47 is based on two steps. In the first one, a state transition graph is created a priori, as described in subsection 4.2.3.1. In the second one, the graph is used in the likelihood scoring of the considered pieces of information, as described in subsection 4.2.3.2. An implementation for axle counters is presented for both steps.

4.2.3.1. State transition graph

One of the meta-information required by the ReCLiC method is the admissible sequences of messages, represented by a state transition graph, as detailed in Section 3.1.2, p. 36. There are two possibilities to obtain this information, either by expert validation or data extraction. Both advantages and disadvantages of the two approaches have been discussed in Section 3.1.3.1, p. 37. They are applied in turn for the case of axle counters below and then compared.

Automatic extraction In the case of axle counters, we automatically extract a state transition graph from the MoTRicS2015 database, analysing the pairs of consecutive messages for all considered sensors and topics. Each newly encountered pair is added as a new edge in the graph; if it has already been observed, its number of occurrences is increased. The number of sections was limited to omit backup tracks or the rarely used ones, resulting in 39 sections out of a to-

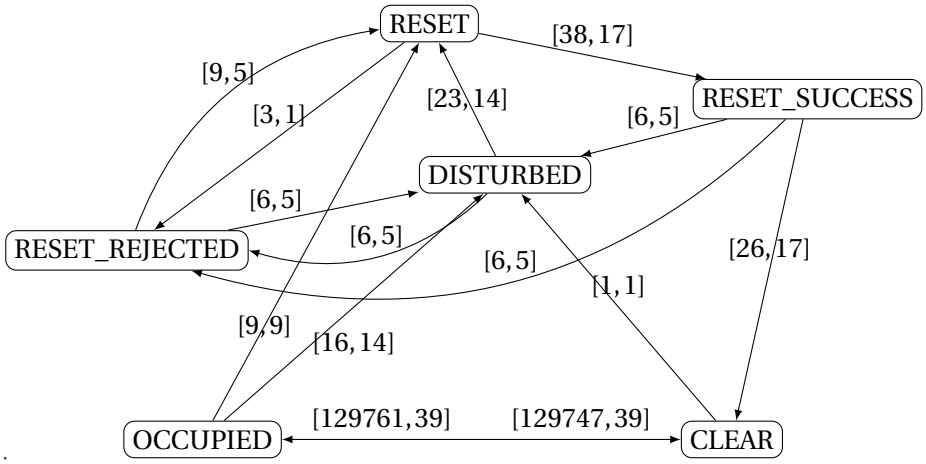


Figure 4.3: State transition graph automatically extracted from MoTRicS2015. Edge labels: [number of occurrences in all sections, number of sections it occurs in], all sections = 39.

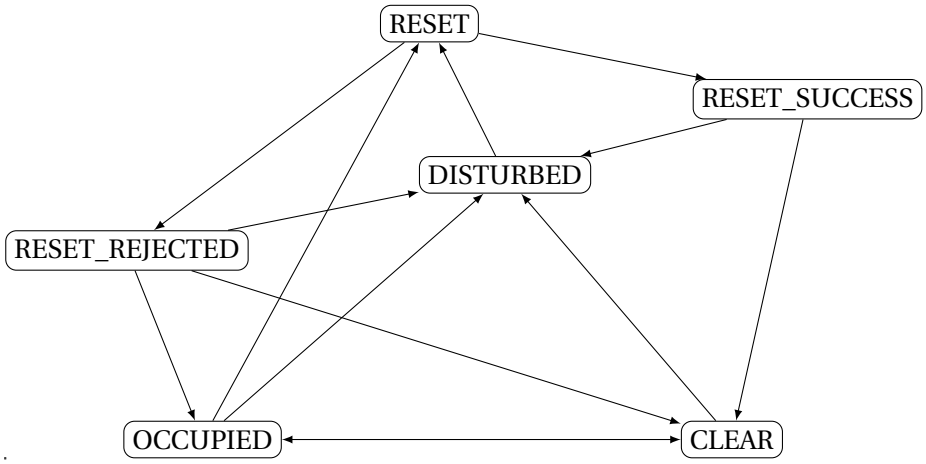


Figure 4.4: State transition graph provided by the Thales Polska expert.

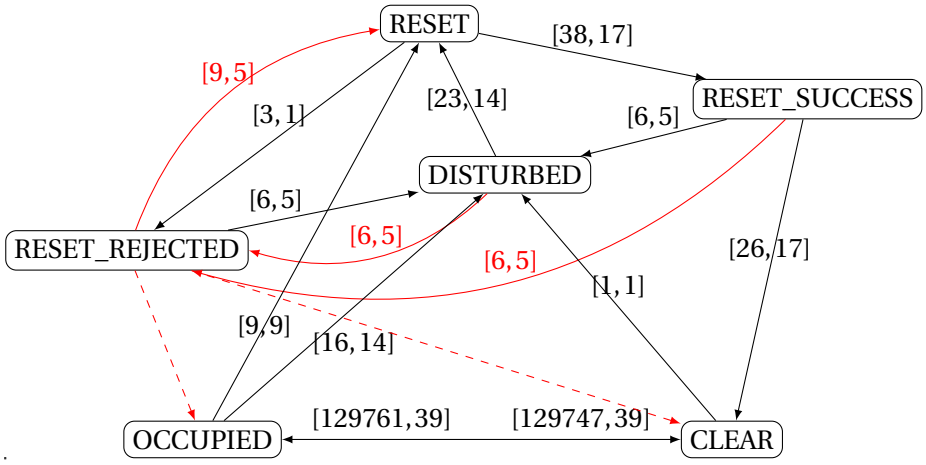


Figure 4.5: Comparison between graphs of Figures 4.3 and 4.4. Edge labels are taken from the data graph. Solid red edged are present in Figure 4.3 but not Figure 4.4, dashed red edges, on the contrary, are present on Figure 4.4 but not Figure 4.3.

tal of 57. Moreover, only the first month of the data was considered, as the remaining dataset can contain quality problems which could add noise in the created state transition graph. The number of considered log entries is approximately 130 000.

All states and transitions encountered in the MoTRicS2015 database are presented in Figure 4.3. Nodes represent all different type of messages that can be produced by axle counters, directed edges between nodes show that these two consecutive messages are present in the database. Additional information is attached in the form of two numbers $[x, y]$ where x shows the total number of occurrences of that transition in the data and y shows the number of sections where at least one of these transitions was present.

Expert definition The second approach for obtaining a state transition graph is to consult with an expert who has deep knowledge about the considered system and devices it uses. In this case, the expert from Thales Polska who provided the MoTRicS2015 dataset was consulted. The resulting graph is shown in Figure 4.4.

Comparison By comparing the two graphs we can see that they are not identical: the state transition graph automatically extracted from the data has some edges missing and some additional ones that are not allowed according to the expert. These differences are highlighted in Figure 4.5 in red: solid edges show transitions that exist in the data-driven graph but are not validated by the expert and dashed edges show transitions that are not observed in the considered data but are defined as possible by the expert.

The problem in the considered graph for axle counters is that the invalid transitions do not stand out neither by the x or y factor: none of the two numbers are neither too low or too high compared to the transitions proposed by the expert. By having both graphs we can already show that some quality problems exist within the considered database as in theory, both graphs should be the same.

For the case of axle counters considered in this thesis, the expert graph is used as the additional transitions provided by the data-driven graph can result from problems in the part of dataset used to create it.

4.2.3.2. Likelihood computation

With the state transition graph available, computing likelihood is straightforward, there is no specificity when considering the case of axle counters as compared to the general case discussed in the previous chapter: in the case of axle counters as well, the previous message of the same sensor and the same topic is extracted and both the previous and the current messages are checked for compatibility with the extracted state transition graph: two messages are compatible if an edge exists from the first one to the second one. The final formula is identical the

one presented in Section 3.5, p. 47:

$$lkl(s,z,m) = \begin{cases} trust(prv(s,z)), & \text{if } prv(s,z) \text{ compatible with } m \\ 1 - trust(prv(s,z)), & \text{otherwise} \end{cases} \tag{4.3}$$

where s is the sensor producing the current piece of information m on topic z, $prv(s,z)$ returns the previous message from sensor s and topic z, $trust(m)$ returns the trust value of message m. Message m_1 is compatible with m_2 only if there is an edge in the state transition graph from m_1 to m_2.

4.2.4. Credibility

The fourth dimension looks for confirmation and integrates messages from other sources. The approach we propose to score credibility is based on spatial confirmation where the distance between the devices plays a major role (see subsection 3.6.2, p. 50).

The general equation (3.4), p. 50 requires to define the expected correlation between the considered sensors, as well as confirming and invalidating messages and an appropriate aggregation. All three aspects are discussed and defined in turn for the case of axle counters in the next subsections.

4.2.4.1. Sets of candidate sources

The first step is to extract a subset of sources that can be correlated with the sensor that produces the evaluated message. In the ReCLiC model, it is achieved by considering a *network of sensors* (see subsection 3.1.3.2, p. 39) where these candidate sources are the ones connected to the considered one.

The proposition described in subsection 3.6.2, p. 50, relies on the geographical distance between devices, where the smaller the distance, the more correlated the messages from these sources are expected to be. This principle indeed applies to the case of axle counters, where there is expected correlated behaviour with sensors that are close to each other. The correlative behaviour can be observed when a train travels a specific route, as the train moves, it sequentially triggers detection points on axle counters along its path without the possibility to skip any (see subsection 4.1.2, p. 63). For instance, when a message *occupied* is produced by an AC with respect to a section, we can assume that one of the neighbour sections reported *occupied* recently and that another neighbouring section will report *occupied* soon as well, as the train has to travel from one way and continue its path. These neighbouring sections can be processed by the same AC or by different AC that are close one to each other. Note that in this discussion, sections are not understood as message topics, as is the case of the other parts of this chapter, but as an indication of sensor proximity.

In the case of axle counters, the *sensor network* is identical to the plan of the considered

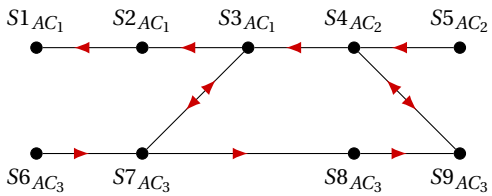


Figure 4.6: A simplified illustrative example of a sensor network graph.

train station where one can find which sensors and sections are close to each other and what is the distance between them. They also provide information regarding the sources that are connected by tracks to handle the situation where two sensors are close to each other but are not correlated (see subsection 3.6.2, p. 50).

The basic approach is to connect all sections to each other and take into account a distance as an indicator of correlation between them. We propose a more tailored approach that connects two sections only if there are tracks joining them. Only direct neighbours are considered which reduces redundancy and simplifies computation: any information from a neighbour is first validated using its neighbours which is validated by its neighbours as well. This creates a chain which includes all sensors starting with the beginning of the train's route to the considered source. In addition, two sets of tracks next to each other represented by two sections are not connected even though there are close to each other.

To summarise, in the proposed implementation, the connection in the network of sensors graph is created if two sections are close to each other. Both source and topic are considered as one AC informs about several sections. No weights for the edges are used. The created graph is a simplified form of the station's plan where the location of sections is shown. An example of such a graph is shown in Figure 4.6 where, for instance, the neighbours of section 3 ($S3_{AC_1}$) are $S2_{AC_1}$, $S4_{AC_2}$ and $S7_{AC_3}$. The exact sensor network is not reproduced due to data confidentiality requirements.

4.2.4.2. Sets of candidate messages

After obtaining the relevant neighbours, the next step is to choose the messages that can serve either as confirmation or invalidation. Since axle counters produce event-type information, the approach presented in Section 3.6.3, p. 51 is applied, identifying specific conditions when a message can confirm or invalidate the evaluated piece of information.

These conditions can be formalised by defining tuples in the form of (s_t, m, w_c) which correspond to a type of sensor, a type of message and considered time window. Since in the considered case, the only type of sensor is axle counter, the s_t value can be omitted: it is the same for all tuples. We propose to consider only the last message of each sensor of all its sections which corresponds to setting $w_c = 1$, where w_c is understood as an absolute number of messages and not a time interval. This last message represents the current state of the considered sections which is the only one relevant to be used as confirmation or invalidation.

As discussed above, the types of messages provided by axle counters we want to score are limited to: *occupied* and *clear*. For each of them, specific conditions need to be defined to extract the messages from correlated sources that can be used as confirmation or invalidation.

In the following paragraphs, we define scenarios for invalidation and confirmation. A scenario is a way to translate a real situation into the combination of messages in the specified order from specific sources.

Invalidation After consulting with the expert from Thales Polska, we propose one scenario which can be treated as invalidation for both messages *occupied* and *clear*. In this scenario, all messages from the correlated sensors must have value *clear* to invalidate the evaluated piece of information. The reason for this depends on the type of message. In the case of *occupied*, one of the neighbours had to report *occupied* as well, since the train had to be present in one of the neighbouring sections in order to travel to the current one: if all neighbours reported *clear*, the current *occupied* is invalidated. Similarly in the case of *clear*, one of the neighbours had to report *occupied*, since the train had to enter one of the neighbouring sections before leaving the evaluated one. Thus, at least one of neighbours has to report *occupied*, all neighbours cannot provide the message *clear* in any of these two cases.

Even though scoring credibility by only considering the above approach to invalidation is sufficient, this type of situation is very rare, leaving credibility to recognise only critical cases. To increase the importance of credibility in the case of axle counters, it is possible to differentiate the remaining possibilities to *confirming* and *neutral* where the latter does not influence the evaluated piece of information in any way. The gain for this additional distinction is obvious as more potentially problematic situations can be recognised, however, the risk is that in some cases information labelled as neutral might unnecessarily decrease credible information.

Confirmation When a sensor provides the information *occupied*, it means that a train entered a specific section. Since this train cannot appear in the middle of tracks, one of the neighbouring sensors should also have reported *occupied* recently. Moreover, since the route of the train is prepared in advanced, one of the other remaining neighbouring sections has to be unoccupied thus providing message *clear*. There is one exception where an additional wagon is attached to an existing train. The wagon and the train occupy neighbour sections, thus the situation where two neighbours report *occupied* is also possible.

A similar scenario can be defined for message *clear* that indicates that a train left a specific section. By leaving the considered section, a train had to appear first in one of the neighbouring ones, i.e. produce *occupied* message by a correlated sensor and section. Moreover, since a train travels in one direction if the current section is *clear*, it indicates that one of the neighbouring section had to report *clear* as well.

	occupied	*clear*
confirmation	$\exists m_1, m_2 \in \mathscr{M}_n$: $m_1 = occ \wedge m_2 = clear$ $\vee$ $m_1 = occ \wedge m_2 = occ$	$\exists m_1, m_2 \in \mathscr{M}_n$: $m_1 = occ \wedge m_2 = clear$
invalidation	$\forall m \in \mathscr{M}_n : m = clear$	$\forall m \in \mathscr{M}_n : m = clear$

Table 4.1: Confirming and invalidating scenarios for two types of messages from axle counters. $\mathscr{M}_n$ represent the set of recent messages output by neighbours of the source of the considered piece of information using sensor network.

For both types of messages, the proposed evaluation does not consider the border case with a single neighbour. Since it is not possible to know whether the train starts at that point or finishes there, all possibilities are equally viable.

All scenarios for confirmation and invalidation are summarised in Table 4.1. Each cell contains the discussed propositions for identifying the required messages from correlated sections. The set $\mathscr{M}_n$ contains the last messages from all n neighbours. We can observe that scenarios for confirmation and invalidation are disjoint. It means that it is not possible to have both invalidation and confirmation at the same time.

Other messages Messages other than *occupied* and *clear* can be considered neither as confirmation nor as invalidation, as thus belong to the neutral category. Indeed, as discussed earlier, *disturbed* is interpreted as an error message, giving an indication about the internal state of a sensor, but not useful to score the messages provided by its neighbours. The *reset*-related messages (*reset*, *reset_ rejected* and *reset_ success*) result from the interaction with the operator, they also give indication about the internal state of a sensor. Therefore when they occur, these messages are considered as neutral messages.

Possible extensions By considering only one type of sensor (AC), the credibility scoring is limited. However, the versatility of the proposed method allows to include easily more confirming or invalidating messages. In particular, it is possible to include additional sensors e.g. lights or point machines which can inform about some contradictions: for instance, if the light is red, then the neighbour AC reporting *occupied* can be invalidated by this source.

4.2.4.3. Aggregation

Choosing the type and complexity of the aggregation operator depends on the available information, as discussed in subsection 3.6.4, p. 52. In the case of AC, since both confirmation and invalidation have additional trust values, we can set the considered complexity level to the one that includes associated quality scores.

Most AC have more than one neighbour and several messages are expected to consider that the message to be score is fully confirmed or fully invalidated. Thus several weights for the

OWA operator need to be set. In the case of invalidation, all messages have to be *clear* in order to fulfil the requirements. There are two possibilities when it happens: either there is a problem with the device and indeed the message should not be produced, or there is a problem with a neighbour producing a wrong message. To decrease the impact of the latter, we set the biggest weight to the message with the lowest trust score: if the message of the neighbour should be trusted, then the invalidation is strong. Otherwise, there is a risk that the problem is with a neighbour rather than the evaluated sensor, thus the invalidation should be weaker.

The weights are proposed as follows $W = [w_1, w_2, w_3, ..., w_n]$, where n is the number of messages from neighbouring sources, $w_n = \frac{1}{2}$ and $w_j = \frac{1}{2\cdot(n-1)}$ for $j \neq n, n > 1$: half of the invalidation score is due to the message with the lowest trust, the other half is redistributed for the rest of messages.

In the case of confirmation, for *occupied* as well as for *clear*, two messages are necessary to fully confirm the evaluated message, they are treated equally. Therefore the OWA weights are set to $[\frac{1}{2}, \frac{1}{2}]$, making the aggregation correspond to the average. If the considered messages are of low quality, then the confirmation is low. This can, for instance, happen if the neighbours do not work correctly, which can indicate that the messages they produce should not be used to confirm the evaluated one.

Final credibility aggregation In subsection 3.6.5, p. 55 a variety of possible aggregations is presented, depending on the required behaviour. In the considered case of axle counters, both confirmation and invalidation are available. In addition, they are "separable": only confirmation (c) or invalidation (i) can have a non-zero value at any time: $c \cdot i = 0$. This feature is a result of the defined disjoint scenarios for both confirmation and invalidation.

The chosen aggregation offers a compromise behaviour between confirmation and invalidation. If neither is present, the credibility score has the middle value of $\frac{1}{2}$. Then the value can be increased by confirmation or decreased by invalidation (see Section 3.6.5, p. 55). The final credibility equation is:

$$cr(s, z, m) = agg_1(c, i) = \frac{c + (1 - i)}{2} \tag{4.4}$$

where c is the confirmation score and i is the invalidation score provided by their respected OWA aggregation in the previous section. As a consequence, in the presence of confirmation, credibility equals $\frac{1}{2} + \frac{c}{2}$ and in the presence of invalidation, credibility equals $\frac{1}{2} + \frac{i}{2}$.

4.2.5. Trust

The four proposed dimensions are finally aggregated into the trust score. This score shows how relevant is the information from an axle counter saying that a train entered the specific section at the specific time (*occupied*) or that a train left the specific section at the specific time (*clear*).

Dimension	**Equation**
Reliability	$r(s,t) = 1 - \frac{e_s}{w}$
Competence	$co(s,z,t) = 1 - \frac{e_{s,z}}{w_{s,z}}$
Likelihood	$lkl(s,z,m) = \begin{cases} trust(prv(s,z)), & \text{if } prv(s,z) \text{ compatible with } m \\ 1 - trust(prv(s,z)), & \text{otherwise} \end{cases}$
Credibility	$cr(s,z,m) = \frac{c + (1-i)}{2}$
Trust	$trust(r,co,lkl,cr) = \alpha \cdot r \cdot co \cdot lkl + (1-\alpha) \cdot cr$

Table 4.2: Implementation of the ReCLiC model for axle counters: summary of the four dimension scoring as well as the final trust score.

In Section 3.7, p. 57 different approaches are considered depending on the user needs: either highlighting any problems detected by either dimension or focusing only on the major breakdowns that trigger more than one dimension at the same time. In the case of axle counters, we recommend the default, compromise, aggregation as it creates a balance between these two extremes.

The final trust aggregation is, therefore, in the form of:

$$trust = \alpha \cdot r \cdot co \cdot lkl + (1-\alpha) \cdot cr \tag{4.5}$$

where reliability, competence and likelihood are combined together and represent the "internal" state of the sensor and the credibility can either increase or decrease this value. The α value is set to 0.75 as the left part combines most of the dimensions and scores both the source and the produced information. However different α values are experimentally studied to illustrate the difference between internal and external scoring in Chapter 5 (see subsection 5.4.3, p. 102). Other possible aggregation operators are studied as well (see subsection 5.4.2, p. 99), to observe their differences and characterise the scenarios where they appear as relevant alternatives.

The final definition of the ReCLiC implementation in the AC case is summarised in Table 4.2. This model is experimentally studied in the next chapter.

4.3. Formal study of trust evolution and its components

This section analyses the trust behaviour by studying formally the equations defining each dimension, as well as the final aggregation into the trust value. By analysing each dimension

separately, we want to emphasise their influence to modify trust over time.

In this section, we consider two successive messages, l_1 and l_2, provided by a given sensor about a given topic, associated to their respective four dimension scoring, (r_1, co_1, lkl_1, cr_1) and (r_2, co_2, lkl_2, cr_2) leading to their respective trust scores $trust(r_1, co_1, lkl_1, cr_1)$ and $trust(r_2, co_2, lkl_2, cr_2)$. We study the difference between these two values, defining $\Delta trust$:

$$\begin{aligned}\Delta(trust) &= trust(r_2, co_2, lkl_2, cr_2) - trust(r_1, co_1, lkl_1, cr_1)\\ &= \alpha \cdot r_2 \cdot co_2 \cdot lkl_2 + (1-\alpha)\cdot cr_2 - (\alpha \cdot r_1 \cdot co_1 \cdot lkl_1 + (1-\alpha)\cdot cr_1)\end{aligned}$$

This difference makes it possible to study the level of decrease or increase of trust. An increase following a decrease period is called *recovery*, it focuses on the ability of the system to get back to fully trusted messages after a decrease has been observed, i.e. after an issue in the quality of the produced messages has occurred.

4.3.1. Reliability analysis

Let us consider the influence of reliability independently of the other dimensions by making co, lkl and cr constant, then:

$$\begin{aligned}\Delta(trust) &= trust(r_2, co, lkl, cr) - trust(r_1, co, lkl, cr)\\ &= \alpha \cdot co \cdot lkl \cdot (r_2 - r_1)\\ &= \alpha \cdot co \cdot lkl \cdot \Delta(r)\end{aligned}$$

i.e. the evolution of trust equals the level of reliability change multiplied by $\alpha \cdot co \cdot lkl$. Therefore, reliability has the biggest impact when competence and likelihood take their maximum value, $co = 1$, $lkl = 1$.

Reliability depends on the *recent* function which returns the w previous entries of the same sensor. Within those w entries, we e denote the number of error entries and m the regular entries, $e + m = w$. When considering the next entry we discard the oldest one. This gives two possibilities: if both entries belong to the same group (error or regular), then the change in reliability is 0, $\Delta(r) = 0$. If they belong to different groups, there are two possibilities as well: either the new entry is an error (m decreases and e increases by 1) or it is a regular message (m increases and e decreases by 1). In these cases, the change in reliability can be a decrease:

$$\Delta(r) = \frac{w-(e+1)}{w} - \frac{w-e}{w} = -\frac{1}{w}$$

or recovery:

$$\Delta(r) = \frac{w-(e-1)}{w} - \frac{w-e}{w} = \frac{1}{w}$$

As a result, it holds that

$$\Delta(r) \in \left\{-\frac{1}{w}, 0, \frac{1}{w}\right\}$$

There is a single level of decrease and recovery, determined by the size of the time window.

Therefore, it holds that

$$\Delta(trust) \in \left\{-\frac{co \cdot lkl}{w}, 0, \frac{co \cdot lkl}{w}\right\}$$

and the trust evolution is bounded by $\frac{1}{w}$ in absolute values. These results shed new light on the w parameter: the time window defines the history horizon to be taken into account, it also has a direct influence on the impact of any error message.

4.3.2. Competence analysis

Let us consider the influence of competence independently of other dimensions by making r, lkl and cr constant, then:

$$\begin{aligned}\Delta(trust) &= trust(r, co_2, lkl, cr) - trust(r, co_1, lkl, cr)\\ &= \alpha \cdot r \cdot lkl \cdot (co_2 - co_1)\\ &= \alpha \cdot r \cdot lkl \cdot \Delta(co)\end{aligned}$$

i.e. the evolution of trust is expressed as the level of competence change multiplied by $\alpha \cdot r \cdot lkl$. Therefore, competence has the biggest impact when reliability and likelihood take their maximum value, $r = 1$, $lkl = 1$.

Similarly to reliability, competence is based on recent error messages. Among w recent messages from all topics of a given sensor, the ones from the considered topic are extracted with their cardinality set as $w_{s,z} \in [0, w]$ (see subsection 4.2.2, p. 67).

Unlike reliability, competence change is difficult to predict and much more different cases need to be considered: it depends whether the first message in the window deals with the considered topic or not, whether it is an error or not, and similarly for the latest message integrated in the window, leading to 16 different configurations. A coarser bound can then be generally established, only stating that $\Delta(co) \in [0, 1]$ and, thus, that the possible trust evolution is: $\Delta(trust) \in [-\alpha \cdot r \cdot lkl, \alpha \cdot r \cdot lkl]$.

4.3.3. Likelihood analysis

To study the influence of likelihood on trust, let us consider r, co and cr as constant, then:

$$\begin{aligned}\Delta(trust) &= trust(r, co, lkl_2, cr) - trust(r, co, lkl_1, cr)\\ &= \alpha \cdot r \cdot co \cdot (lkl_2 - lkl_1)\\ &= \alpha \cdot r \cdot co \cdot \Delta(lkl)\end{aligned}$$

Again, we can observe that the trust evolution is proportional to likelihood, with multiplicative factor $\alpha \cdot r \cdot co$, which means that the biggest decrease happens with r and co at maximum value, $r = 1$, $co = 1$.

Now the likelihood values of the messages l_1 and l_2 are not directly dependent one on the other, contrary to their reliabilities and competences. However, one can highlight some interesting properties. Let us introduce a third message l_0, produced even before l_1, so that we have the three consecutive messages l_0, l_1 and l_2 provided by the same sensor about the same topic. Let us assume that they have the same reliability, competence and credibility.

We then focus on two cases, when both transitions, from l_0 to l_1 and from l_1 to l_2, are valid according to the considered state transition graph and when both are invalid (see Section 3.5, p. 47).

In the first case, according to the ReCLiC model, $lkl(l_1) = trust(l_0)$ and $lkl(l_2) = trust(l_1)$, which leads to

$$\begin{aligned} \Delta_2(trust) &= tr(l_2) - tr(l_1) = \alpha \cdot r \cdot co \cdot (lkl_2 - lkl_1) \\ &= \alpha \cdot r \cdot co \cdot (tr(l_1) - tr(l_0)) = \alpha \cdot r \cdot co \cdot \Delta_1(trust) \end{aligned}$$

Thus we can observe that if the information flow appears as likely, then trust continues its previous decrease or recovery multiplied by $\alpha \cdot r \cdot co$. Since $\alpha \in [0,1]$, $r \in [0,1]$ and $co \in [0,1]$, it makes each following decrease or recovery equal or slower: $|\Delta_n(trust)| \leqslant |\Delta_{n-1}(trust)|$.

On the other hand, when the two consecutive pieces of information are not compatible with the model, $lkl(l_1) = 1 - trust(l_0)$ and $lkl(l_2) = 1 - trust(l_1)$, leading to

$$\Delta_2(trust) = -\alpha \cdot r \cdot co \cdot \Delta_1(trust)$$

We can see that in a case where the consecutive pieces of information are considered as unlikely, likelihood reverses the trend of trust, changing decreasing to recovery and vice versa. This observation shows that in continuous quality problems detected by likelihood, a fluctuation behaviour can be observed, where the trust evolution can consist in alternated decreases and increases. Again, since it is multiplied by $\alpha \cdot r \cdot co$, where $\alpha \in [0,1]$, $r \in [0,1]$ and $co \in [0,1]$, $|\Delta_2(trust)| \leqslant |\Delta_1(trust)|$ which creates the oscillation of equal or decreasing magnitude.

4.3.4. Credibility analysis

The credibility dimension is defined independently from the previous three dimensions and its scoring is based on the comparison of the considered piece of information with the ones provided by different sensors. When considering r, co and lkl as constant, it holds that

$$\begin{aligned} \Delta(trust) &= trust(r, co, lkl, cr_2) - trust(r, co, lkl, cr_1) \\ &= (1-\alpha) \cdot (cr_2 - cr_1) \\ &= (1-\alpha) \cdot \Delta(cr) \end{aligned}$$

We can notice that the actual values of the constant r, co and lkl are irrelevant when analysing the impact of credibility on trust difference. The evolution of trust is proportional to credibility, with multiplicative factor $(1-\alpha)$. However, there is no explicit correlation between the previous credibility value and the current one, nor with any previous dimension. Therefore we limit $\Delta(cr)$ only by credibility value itself which is in $[0,1]$ range. Then, by including $(1-\alpha)$ factor, the possible trust evolution: $\Delta(trust) \in [\alpha-1, 1-\alpha]$.

4.4. Summary

This chapter discussed in details the implementation of the ReCLiC model to a real case, showing how to adapt the generic ReCLiC definitions proposed in the previous chapter to a very concrete case, that of the specific sensor called axle counter used in the railway signalling domain. After presenting this domain and its monitoring devices, the chapter described the available real dataset, MoTRicS2015. It then discussed, in turn, each of the four dimensions ReCLiC relies on, among others showing how the required easy-to-access meta-information can be automatically derived from the data. It thus provides operational formulae that make it possible to compute reliability, competence, likelihood, credibility, and to derive the trust score to be attached to each information provided by axle counters. It thus allows to increase the monitoring of a train position on tracks and to detect potential quality issues that may lead to the intervention of human operators. The chapter also provides a formal study of the trust evolution, analytically examining the effect of each dimension separately, to characterise the behaviour of the proposed model.

5. Experimental study on realistic simulated data

This chapter presents the experimental study of the information scoring model proposed in the previous chapter, performed on realistic simulated data: it first details the motivation for using such simulated data in Section 5.1, it then presents in Section 5.2 the four scenarios we propose for generating simulated data from real ones. Section 5.3 discusses the results obtained with the ReCLiC model summarised in Table 4.2, p. 76, using the default parameters ($w = 20$ and $\alpha = 0.75$) that allow to validate this choice. Section 5.4 experimentally studies on these data the role of the ReCLiC parameters.

5.1. Motivation

Before applying a new model to real data, this model needs to be validated to assess its relevance and validity. This is a crucial task, whose results may significantly impact any decision-making system. Effective ways of validating a model usually depend on its specificity as each proposed model can consider different dimensions or work with different types of data. Moreover, the successful validation of a model on data from one domain may not be relevant in other domains. Thus, it is important, especially in quality scoring models, to test the considered model on the data it will be used on.

The most common approach to validation is to assess the accuracy of the outcome with respect to reality. The difference between the two shows how good the considered model is. This approach highly depends on ground truth and is applied in many fields like classification, regression, clustering and information scoring as well, in the case where additional information about quality is attached to each piece of information. However, labelled data is often costly or impossible to obtain, especially for trust scoring models where it is difficult to attach an exact expected trust value to each message.

To mitigate this problem, an experimental study on simulated data is proposed: it consists in modifying the original dataset by introducing problematic data. This approach allows to create a controlled test environment, considering the original dataset as ground truth to evaluate the results and especially trust score decrease and recovery, i.e. trust back increase to its nominal level.

Such simulated datasets allow to control the type of introduced quality problems, as well as their intensity and distribution. They allow to cover a large spectrum of problems which include rarely observed issues and study how the evaluated model behaves. These scenarios need to be realistic as they should simulate real problems which can be encountered with the considered type of sensors. The multiplicity of the proposed scenarios can greatly improve the trust level interpretability for the end-user.

Section 5.2 describes the four different scenarios we propose. Each of them aims at evaluating a specific problematic situation that can be encountered. The ReCLiC model as summarised in Table 4.2, p. 76, is tested on the generated data. Section 5.3 comments the observed values for the global trust score, as well as for each of the four dimensions ReCLiC relies on, considering for all of them their temporal evolution and respective impacts on one another. Section 5.4 analyses the influence of the parameters the ReCLiC model depends on (namely the temporal window w and the aggregation weight α), studying the result changes when the parameter values change. Due to the specificity of existing information scoring models as described in Chapter 2, it is not possible to apply them to the simulated data and to perform a comparison with the ReCLiC results.

5.2. Proposed scenarios for the generation of realistic simulated data

After giving an overview of the proposed scenarios to generate the simulated data, this section describes each of them in turn, providing in each case a description and a discussion of its specific aim.

5.2.1. Overview

As mentioned in the previous section, the principle of simulating realistic simulated data consists in introducing noise in real data, to introduce some problematic log entries in the available logs. For the data to be relevant for the result analysis, it is necessary that the introduced noise correspond to the only problematic log entries in the data. Therefore, all the simulations are based on a subpart of the original MoTRicS2015 data (as described in Section 4.1, p. 61) for which the ReCLiC model outputs a constant trust value of 1: this indeed indicates that no issue is encountered in this part of the data and that all log entries can be trusted. Among others, for instance, this part of data does not contain any *disturbed* message that would decrease reliability and thus trust. More precisely, this original data corresponds to the dates between 2015-03-08 and 2015-03-12 from one station. In the following experiments, axle counter *AC4* is considered, it manages 10 sections, among which section *JT1258* is used. The chosen section is the most common one, it is in particular frequently used as opposed to

backup tracks.

We propose four types of noise, that realistically simulate different behaviours of faulty sensors. They can be categorised into two types: in the first one, existing log entries are modified by changing the type of information they carry. This approach allows to keep the original structure of the produced log entries, including the time in relation to the previous entries. In the second type, noise is introduced by randomly injecting new entries: this approach allows to simulate a different situation where a sensor unexpectedly produces information.

Besides these two categories, two possibilities for generating noise for a given sensor can be considered: noise can affect entries related to all topics associated to the sensor or only entries related to a single, fixed, topic. The first approach makes it possible to observe a rich evolution of reliability and its influence on trust for all entries output by the sensor. On the other hand, as illustrated in the preliminary experiment described in Section 5.3.2, p. 86, this approach makes it difficult to analyse the evolution of the other dimensions, that all depend on a single topic. Therefore, for most considered experiments, noise is only introduced for a single sensor and a single topic associated with the sensor.

5.2.2. Uniform noise

In the first scenario, randomly chosen log entries in the original data are modified by changing the type of information they carry to a different one, also produced by this device, they modify the field *l.msg* for an entry *l*.

For the case of axle counters considered here, the two most frequent messages are considered, *clear* and *occupied*. The proposed modifications are: *clear*, *occupied* and *disturbed* with equal probability. The message *disturbed* is included as it is crucial in reporting errors for AC which are used to score reliability and competence in the ReCLiC model (see Chapter 4).

These modifications are performed for randomly chosen log entries, uniformly distributed in the entire dataset. To observe how this kind of individual issues impact trust decrease and recovery, we introduce the variable η that corresponds to the probability of a log entry being modified. In the case of uniform noise, $\eta = 3\%$ of the log entries are modified.

This type of simulation corresponds to a situation where messages are corrupted due to a sensor malfunction or communication interferences. It results in modified messages and possibly unjustified error ones.

5.2.3. Burst noise

The burst noise scenario aims at introducing noise only within a small time window, containing 30 log entries, and not across the whole considered period, but with a high probability of altering a log entry $\eta = 90\%$. This results in many disruptions in a short period of time. As in

the previous scenario, the two most frequent messages are considered, *clear* and *occupied*, and the proposed modifications are: *clear*, *occupied* and *disturbed* with equal probability.

The aim of this scenario is to simulate a broken device which produces random log entries. It allows to observe how the evaluated model behaves when encountering a majority of problematic log entries close to each other and how the trust value recovers when the noisy period ends.

This scenario is the first step towards a study of a critical threshold for the parameter of the uniform noise, i.e. a level of noise above which the whole system would break down. A full study would include testing with many increasing η values, it goes beyond the aim of the experimental study performed in this chapter and it is left as future work.

5.2.4. Random message injection

In this scenario, a second approach to simulate problems is considered which randomly inserts additional entries to the log file. Two attributes of the new log entry are randomised: date and time and the type of information it carries, i.e. its message. More formally randomised the attribute *l.date* and *l.msg*. The injected messages are: *clear*, *occupied* and *disturbed* with equal probability. Noise injection is performed only for one sensor and one topic. The injection probability is $\eta = 3\%$.

This scenario aims at simulating a device which is not only triggered by the event it observes, i.e. due to current overload instead of a train passage for AC. It can be attributed to a device malfunction or a malicious intent of an external agent. The main difference with the previous scenarios is that the injected message is not part of any sequence, neither an internal one with respect to the considered sensor nor an external one with respect to multiple sources, as should normally be observed in the case of multiple correlated sensors.

5.2.5. Non-existent message injection

This scenario is similar to the previous one, it also considers injecting additional log entries to the existing database, at random dates. However, in this case, the type of message is not one produced by the evaluated sensor: the information is random, it does not correspond to the type of messages produced by this device. The exact message of an injected log entry is not important as it differs from the current types of messages. In a simulation tool, three different messages are introduced: *occuppied*, *clearrr* and *dissturbed*, all imitate corruption of the log file.

This approach can simulate data transfer disturbances or malfunctions which alter the original information. It can also be a result of database corruption. As in the previous case, the injection probability is set to $\eta = 3\%$.

5.3. Illustrative results

In this section, the results of applying the ReCLiC model to the case of AC are presented. The model is the one summarised in Table 4.2, p. 76 with its default parameters: time window for reliability and competence $w = 20$, state transition graph for likelihood as given in Figure 4.4, p. 69, the network of sensors considered for credibility is omitted for confidentiality reasons, aggregation parameter for trust $\alpha = 0.75$.

The four scenarios discussed in the previous section are considered. For each scenario, the results are visually studied by plotting the trust score of each entry over time. In particular, different types of trust decrease are analysed as well as the following trust recovery which is how trust increase back to its nominal level. In addition, the evolution of the four dimensions ReCLiC relies on is studied.

First, the considered visualisation for the results is presented in Section 5.3.1. Then the results for each of the proposed scenarios are presented and analysed in Sections 5.3.2 to 5.3.6.

5.3.1. Considered visualisation

All figures in the chapter have the same form: the x-axis represents time, measured as the log entry number, the y-axis represents the considered dimension (reliability, competence, likelihood, credibility or trust). This is a simplified approach to illustrate trust values for a single sensor: the exact time is not represented since the time between two messages can be a few seconds or a few hours. Because of that, the graph based on time is not clear and it only illustrates a small part of the considered information. As discussed in subsection 5.2.1, during the considered period, no sensor provides low-quality message in the original database to exclude the possibility of any unwanted interference.

Each modified entry is represented by a vertical line, whose colour indicates detail about the change:

- blue: *occupied* changed to *clear*
- orange: *occupied* changed to *disturbed*
- violet: *clear* changed to *occupied*
- green: *clear* changed to *disturbed*

In the second type of noise where the new log entry is created and injected to the dataset, each injected log entry is also highlighted by a vertical line with the following colour code:

- blue: injecting *clear*
- violet: injecting *occupied*

– green: injecting *disturbed*

– black: injecting non-existing message

5.3.2. Uniform noise applied to all topics from one sensor

As discussed in Section 5.2 there are two ways to introduce noise for a sensor. It can be applied to all its topics or limited to one. Both approaches are considered in this section and the following one to illustrate their advantages and disadvantages.

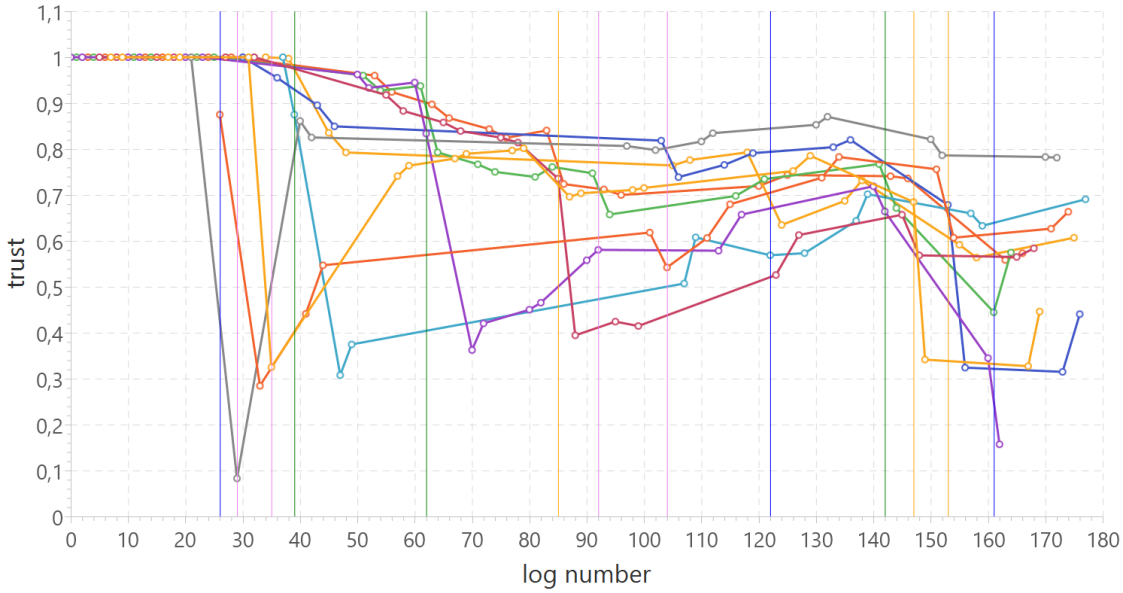
In the first case, messages from all topics of one sensor are considered in one graph to plot their trust and reliability values. The results are shown in Figures 5.1a and 5.1b. Each topic is represented in a different colour and the position of each message is highlighted by a small circle. Actually, all topics should be plotted using the same colour and the circles should be connected only depending on the date of the corresponding messages, but this representation would make the graphic even less legible.

In Figure 5.1a the trust evolution is illustrated. At this point, we can observe that it is very difficult to interpret the individual evolution for each topic. In this graph, it is possible to observe how trust changes for one topic when encountering a simulated entry as well as the following trust changes for other topics. However, the analysis is difficult due to information clutter of multiple curves in one graph. In addition, by simulating noise for all topics, additional complexity is introduced: trust decreases not only because of encountering a simulated entry but also because other topics can have low trust. This makes analysing ReCLiC output more difficult.

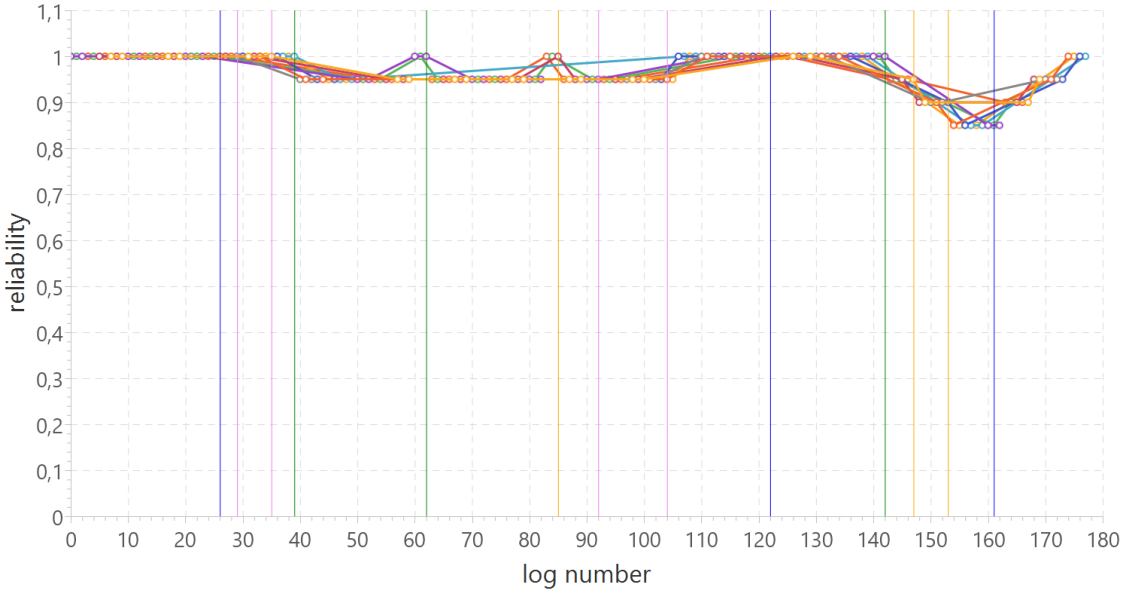
However, this type of visualisation is valuable when analysing the reliability evolution which is illustrated in Figure 5.1b. Since its definition is based on information produced by the considered sensor, that includes entries from all topics. The time window is set to 20 last log entries, thus we can observe that after introducing an error message (coloured as green or orange) the next 20 entries from any topic are decreased equally, regardless whether this topic produced this error or not, e.g. entries 40-60, 62-82 or 85-105. In addition, when within this 20-entry window another error message occurs, we can observe a stair-like behaviour where reliability is decreased again by a constant level and recovers similarly, e.g. entries 142-172. This behaviour has been discussed in the formal analysis of reliability, in subsection 4.3.1, p. 77. The presented figure shows that if only one topic provides error messages, the overall reliability is not highly reduced. Only if most of the topics start producing errors, the reliability gets low, indicating that the entire sensor malfunctions.

5.3.3. Uniform noise applied to a single topic

The trust evolution for a uniform noise applied to a single topic is shown in Figure 5.2a. This approach has several important advantages as compared to the previous one. First and most



(a) Trust



(b) Reliability

Figure 5.1: Uniform noise distribution on all topics of one sensor

importantly, the analysis of the trust evolution is easier as there is less interference from other topics decreasing the trust value. Second, since only messages about one topic are considered, more can be displayed on a graph to study the evolution more thoroughly. Finally having a single topic displayed makes it easier to follow the changes. All these aspects allow to analyse the individual decreases and recoveries of the trust value which are discussed in this section.

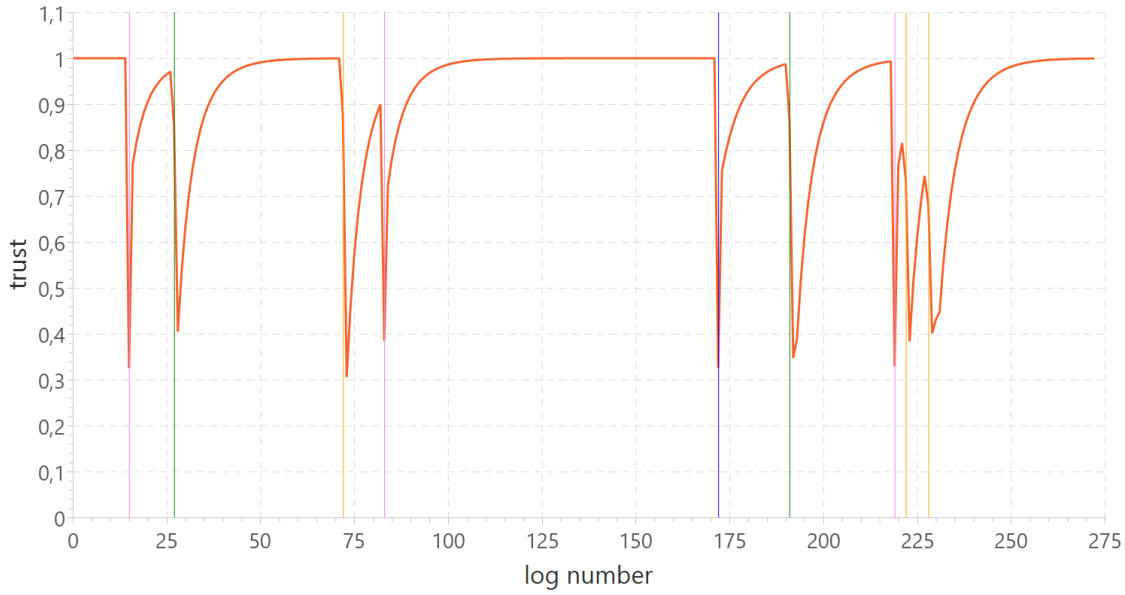
The first important observation is that all non-error simulated entries (shown by the blue and violet vertical lines) indeed have a decreased trust value: this expected result is an important first step in the validation of the proposed ReCLiC model that is able to identify these noisy log entries and to assign them a low quality. A clear example of this decrease is provided by entries 15 or 172 falling from 1 to 0.33. Both illustrate the level of single decrease for non-error noise. Furthermore, they also show the slow recovery which takes place afterwards where trust increases, but not instantly, back to its maximum value. The uninterrupted recovery covers a window of at least 15 entries. As discussed in Section 2.3, p. 33 this behaviour is desired: the trust evolution patterns are expected to be asymmetrical, with fast decreases and slow recoveries. For other non-error entries, this asymmetrical behaviour also holds with slightly lower decrease, which can be explained by the other dimensions, as discussed below.

In the case of error messages (orange and green vertical lines), we can observe that a single disturbance also causes a significant trust decrease with the major difference that this decrease applies to the next entry and not instantly, e.g. entries 73 or 193. Indeed, after such an error message, a correction process is supposed to take place, involving other messages required before the state is again allowed to be *clear* or *occupied* (see the state transition graph, Figure 4.4, p. 69). The absence of these messages thus decreases the likelihood value of the next entry. In addition, such error messages decrease reliability and competence of the next entries. All these reasons explain the reduced trust value, that also corresponds to expected behaviour.

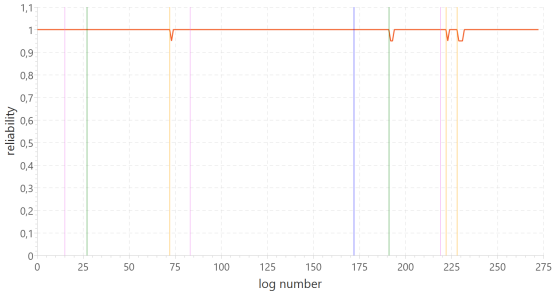
Figure 5.2b illustrates the evolution of reliability. As expected and mentioned above, reliability decreases only after the simulated error-type entries (orange and green lines). The value then remains low for the next 20 entries corresponding to the value chosen for the time window parameter. However, since the figure only shows the log entries corresponding to the chosen topic, the total number of log output by the considered sensor is not displayed. This explains why the duration of the reliability decrease appears to vary: it depends on the number of entries output by the considered sensor but concerning other topics. Note that this comment exploits the fact that the other topics do not produce error messages, due to the considered simulation process and the original data characteristics (see Section 5.2, p. 82).

This simulation, performed with a 3% noise level, does not illustrate the case where 2 error entries are encountered within the considered temporal window (see next section, p. 91 for such a case), the reliability thus only takes two values, 1 and $1 - \frac{1}{w} = 1 - \frac{1}{20} = 0.95$ as discussed in the formal analysis of reliability (see subsection 4.3.1, p. 77).

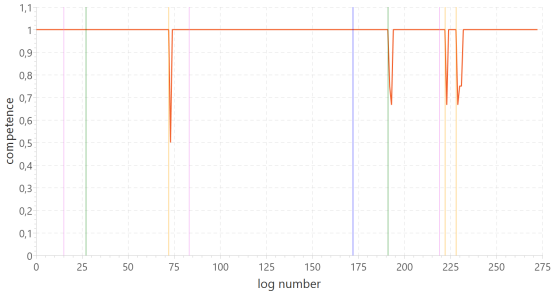
Figure 5.2c shows the evolution of competence. Competence, unlike reliability, depends on



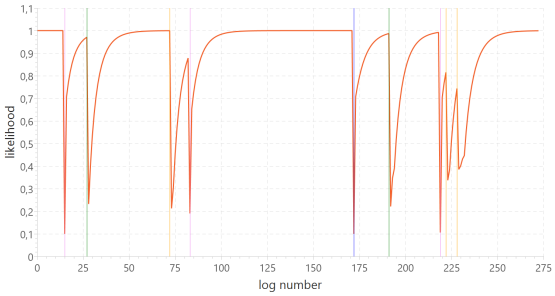
(a) Trust



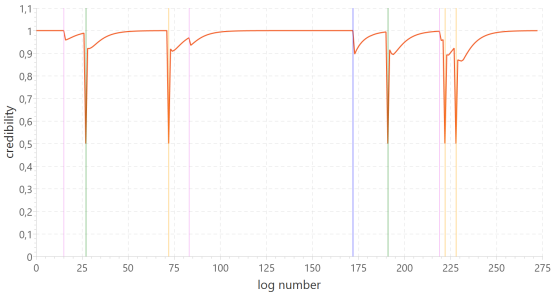
(b) Reliability



(c) Competence



(d) Likelihood



(e) Credibility

Figure 5.2: Uniform noise distribution

the topic of the message. From the last 20 messages considered by reliability, the ones from the evaluated topic are extracted and analysed. It means that fewer entries are considered for competence scoring which results in larger decreases and faster recoveries in competence value compared to reliability. The level of decrease, as well as the later recovery, depend on the activity of the considered topic with respect to the whole activity of the sensor. For instance, in entry 73 the decrease is 0.5 because only two messages are considered, one being an error. Both for this message and for entry 223 the recovery is quick and much quicker than for reliability: this results from numerous entries provided by other topics between this message and the next one where the error message is no longer in the considered time window. In this graph, the time between two log entries is not visualised, however, it can be observed indirectly. For instance, as opposed to the previous example, entry 228 shows no imminent recovery to 1 but stays lower (0.75) one more message, which means that two following messages after the error are within a short time window.

Figure 5.2d shows the evolution of likelihood. We can observe that its shape is highly similar to that of trust, with exactly one timestamp delay: the simulated entries appear to be usually compatible with the previous messages, making the likelihood equal to the trust value of the previous message (see the likelihood definition in Equation 3.3, p. 48). However, this does not hold at the position of a simulated log entry. We can distinguish two behaviours, depending on whether the entry is an error or not: for non-error messages, the likelihood value decreases significantly making it a key dimension in problem recognition for this noise scenario, e.g. entries 15, 83, 172 and 119. This is due to the fact that the simulated entry is most often not compatible with the previous entry, which is highlighted by the state transition graph. The different decreases observed in the figure result from different trust values of the previous log entry (see the formal analysis in subsection 4.3.3, p. 78). In the case of error messages, the likelihood value is decreased not for the simulated message but the following one. In part, this results from reliability and competence considering error entries to decrease trust. In addition, as discussed above, likelihood recognises that the recovery needs to contain a sequence of the messages (see the state transition graph in Figure 4.4, p. 69) and since it is not the case, likelihood is decreased, e.g. entries 27, 72 or 192.

The credibility evolution is illustrated in Figure 5.2e. We can observe a significant decrease only in the case of errors (entries 27, 72, 191, 222 and 228) and their value is 0.5. This behaviour is expected since neither confirmation nor invalidation is defined for this type of log entries, thus credibility takes the middle value (see Section 3.6.3, p. 51). When considering non-error messages, we can observe that the credibility value stays high, it decreases for the next log entry. This behaviour can be explained as follows: the non-error simulated entries are either *occupied* or *clear*, for which the definition of confirmation and invalidation (see Section 4.2.4.2, p. 72) are such that two messages need to be recently produced by two neighbours reporting *occupied* and *clear*. Since in this scenario, the sequences of entries produced due to a train passage

are preserved, the messages required for confirmation are also preserved, even though they are in the wrong order. This results in a high score for confirmation and thus for credibility. On the contrary, for the next log entry, the credibility is slightly decreased. This results from disrupting the confirmation chain for the neighbouring sources, thus decreasing trust for their messages. These messages are then considered by the evaluated sensor and their low trust impacts credibility.

5.3.4. Burst noise

In the second scenario, noise introduction is limited to a small time window which corresponds to approximately 30 log entries. In that window, 90% of the entries are noisy, using the approach that modifies the carried information. This results in a short simulated breakdown that should be easily noticeable by any trust scoring model.

The results of applying the ReCLiC model to this type of noise are illustrated in Figures 5.3a-5.3e. The general overview of trust evolution is presented in Figure 5.3a and indeed shows lowered trust value throughout the entire simulated breakdown. Because of the intensity of this noise, trust does not have time to recover for the unmodified entries within the time window, thus they also have low trust. ReCLiC is able to identify a period of breakdown that casts doubt on all the produced messages in that window which later need some time to recover before again fully trusting the produced messages. It successfully captures the trust to be put in the log entries in a dynamic adaptive way with the desired behaviour.

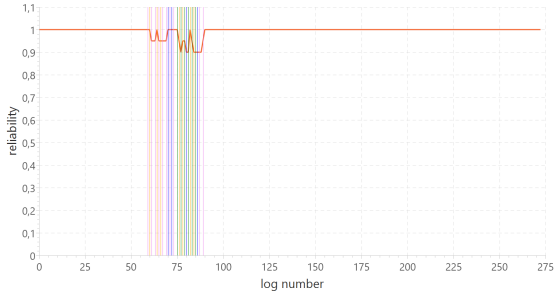
The level of trust varies throughout the noise, it never decreases to 0 nor does it remain constant. In the considered period, there is a mix of error and non-error messages, which impacts all the considered dimensions, trust aggregation of constantly changing scores of each dimension results in various trust levels through the breakdown.

The recovery process of a trust value after the simulated noise proceeds as expected, similarly to a single noise presented in the previous scenario. We can observe some recovery within the noise period, where the trust value is above 0.5 which can be correlated with the high competence score.

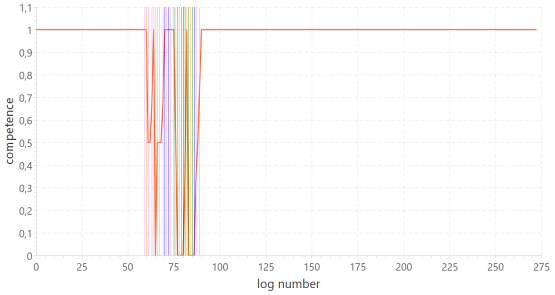
Figure 5.3b illustrates the evolution of reliability. As expected, decreases take place after each error message. One may consider that the decreases are not really significant, as compared to the proportion of error entries, but the latter is actually not so high when compared to the total number of entries produced by the considered sensor in this time window. This is an illustration that reliability is based on all messages from the considered sensor and not only the considered topic. Thus, if error messages are limited to a single topic, the reliability value remains high, showing that information from other topics can still be reliable.

The competence evolution is illustrated in Figure 5.3c. Even though its definition is close to that of reliability, their values differ significantly due to their different normalisation factors: burst noise provides a good illustration of the difference between reliability and competence.

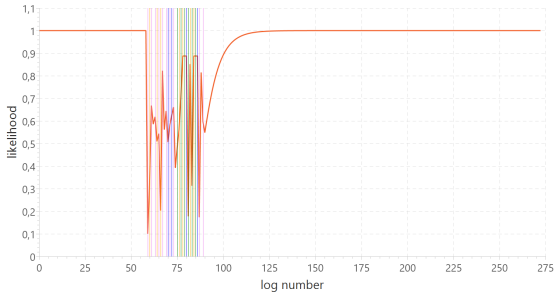
(a) Trust



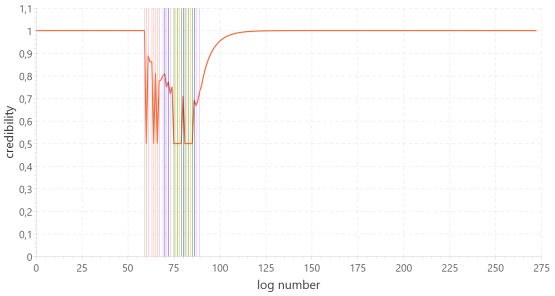
(b) Reliability



(c) Competence



(d) Likelihood



(e) Credibility

Figure 5.3: Burst noise

Whereas reliability remains high throughout the noisy period, competence can decrease significantly when encountering multiple error messages in a short time window. As observed in Figure 5.3c, competence reaches the lowest value 0 multiple times and it is one of the key dimensions in recognising this type of noise.

Figure 5.3d shows the likelihood evolution. We can observe a major variability in the likelihood values throughout the noisy period, that oscillate between high and low levels. This behaviour can probably be explained by the fact that in noise that dense, two following messages might still be randomly compatible with the state transition graph.

The credibility evolution is presented in Figure 5.3e. As in the previous scenario, invalidation is not recognised for any of the simulated entries, thus the credibility value never decreases below 0.5. Moreover, credibility values for error messages remain always at 0.5 as they can be neither confirmed or invalidated (see subsection 4.2.4.2, p. 72). However, we can observe some decreases in the noisy period for non-error messages. This shows that, in the dense noise scenario, credibility can recognise problematic entries. In addition, the level of credibility decreases further when the noise period is longer, leading to a decreasing trend which should be further studied.

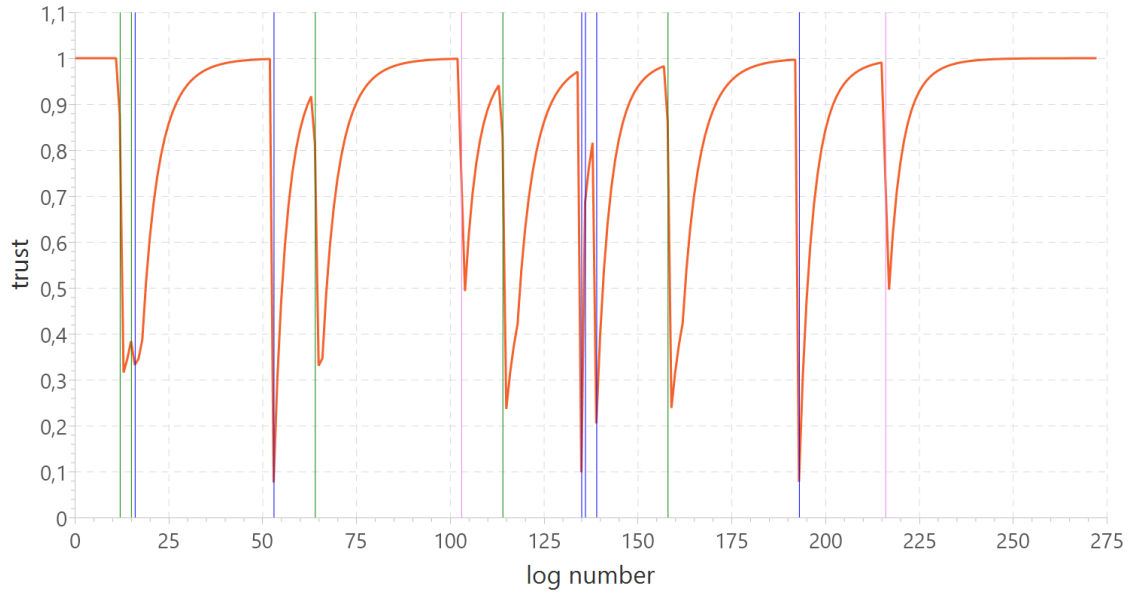
5.3.5. Random message injection

As detailed in Section 5.2, p. 82, another scenario is based on injecting noise instead of changing the existing messages: log entries are randomly created and added to the reference database. The results of applying the ReCLiC model to such a simulated database are shown in Figures 5.4a-5.4e.

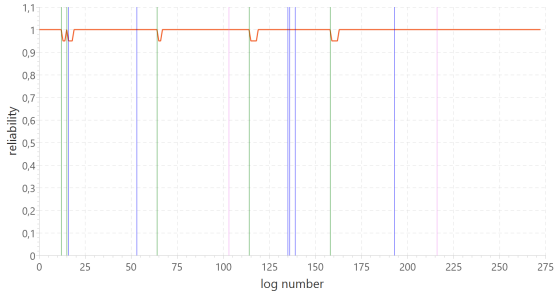
Figure 5.4a presents the trust evolution for this scenario: one can observe a significant decrease for all simulated entries, sometimes with a delay of one entry. However, for each type of noise, the trust value decreases differently: the main difference is between two non-error messages, where trust decreases significantly more in the case of message *clear* (blue vertical lines) to 0.1 than *occupied* (violet vertical lines) to 0.5. For the latter, trust decreases in two steps: it first obtains a lower value for the simulated entry, but it further decreases for the following one. This behaviour is commented below when observing the likelihood evolution.

The exact trust level varies not only for the different types of noise. We can observe that various distributions of noise affect the level of trust. For instance, most often the injection of *clear* (blue line) leads to a trust level decreases to 0.1 point; however, in denser noise period, the trust value decreases less. This behaviour can, for instance, be observed for entry 16 or 136 when trust reaches 0.34 or 0.7 respectively. It shows that the trust level of the second of two consecutive simulated entries can also be low. It shows the effectiveness of ReCLiC model in such difficult situations.

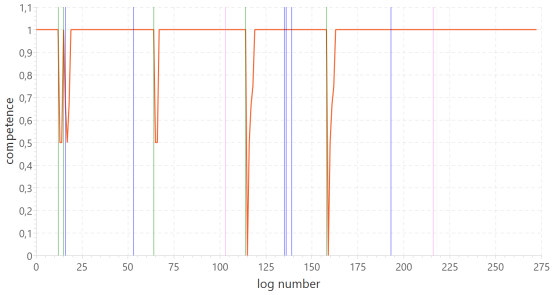
Figure 5.4b illustrates the evolution of reliability. As with the previous scenarios, reliability decreases after error messages and the duration of the decreased reliability value varies



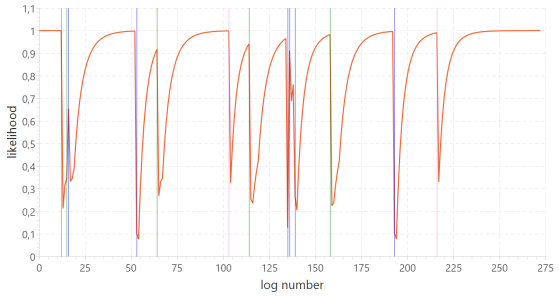
(a) Trust



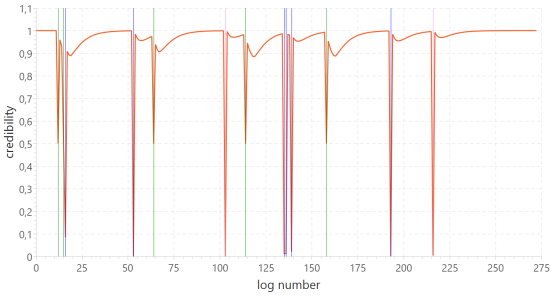
(b) Reliability



(c) Competence



(d) Likelihood



(e) Credibility

Figure 5.4: Random message injection

depending on the total amount of messages produced by the sensor including the messages applying to other topics. The in-depth analysis performed in the previous sections for the previous scenarios also applies to this case.

A similar situation is observed for competence, illustrated in Figure 5.4c: its value decreases after an error message and the decrease is more significant than that of reliability. However, this case differs from the previous scenarios as the messages are randomly injected and there exist no sequences with other entries that are normally present at a similar time. This results in a situation where no other messages can be observed for the same topic except the error one. It increases a chance for lack of competence decrease or a significant one as the number of messages within this time window is small. Because of that, we can observe the competence value as low as 0 or 0.5 or, on the contrary, not decreased at all.

Figure 5.4d illustrates the likelihood evolution. Several interesting observations can be made in this case. First, the recognition of non-error messages differs drastically: it is high for the injection *clear*, but, for *occupied*, the likelihood value decreases only after the injected entry. This delay explains the differences observed for trust scored for two non-error messages (blue and violet) which were recognized differently. This situation has an easy explanation: in this type of noise injection usually the previous message is *clear* as it is the message that ends most sequences, including the most popular one where the train has left the section. If we consider injecting *occupied* afterwards likelihood sees it as a viable transition. Something that is not the case when two consecutive *clear* messages are observed.

Another interesting situation relates to the injected *disturbed* messages (green line) for which the likelihood value decreases for the next entry but not at the time of the injected entry. As mentioned in the previous scenarios, this is due to the fact that, after a *disturbed* message, a recovery procedure is expected but is not observed. As a consequence, the ReCLiC model indicates that the following message should be considered with caution.

Figure 5.4e illustrates the credibility evolution. We can observe that its representation significantly differs from the previous scenarios. In this case, credibility highly decreases for each non-error noisy messages. This decrease is based on observed invalidation as the injected noise does not correlate with any activity of the neighbours. In the case of errors, the situation is similar to the previous scenarios: the message is neither confirmed nor invalidated, thus its credibility score remains at 0.5. We can observe another interesting behaviour after each simulated entry where credibility recovers and then decreases again in a lesser degree. This behaviour can also be observed in Figure 5.2e and it happens due to trust propagation between neighbours (see subsection 6.2.3, p.115): a decrease in trust affects messages of neighbours which then affect the current source and its messages again but in a much lesser degree.

5.3.6. Non-existent message injection

In the fourth scenario, new messages are created which are actually not produced by the considered sensor. These messages are injected into the original database and the ReCLiC model is used to assess the trust associated with the whole log. The results are illustrated in Figures 5.5a-5.5e.

The trust evolution, represented in Figure 5.5a, presents two situations: the first part of the chart, where noise is somewhat denser, from entries 1 to 30, and the second one, after entry 30, where noise is distributed more sparsely. In both cases, the simulated messages have a significant decrease in trust. In the second part, each injected message triggers the same decrease and the same recovery time, see for instance entries 71, 140 or 185 where trust decreases to 0.2 and recovers in 20 entries. For the denser noise, the decrease depends on the previous trust value and the trust level does not have time to recover and remains only slightly over 0.5.

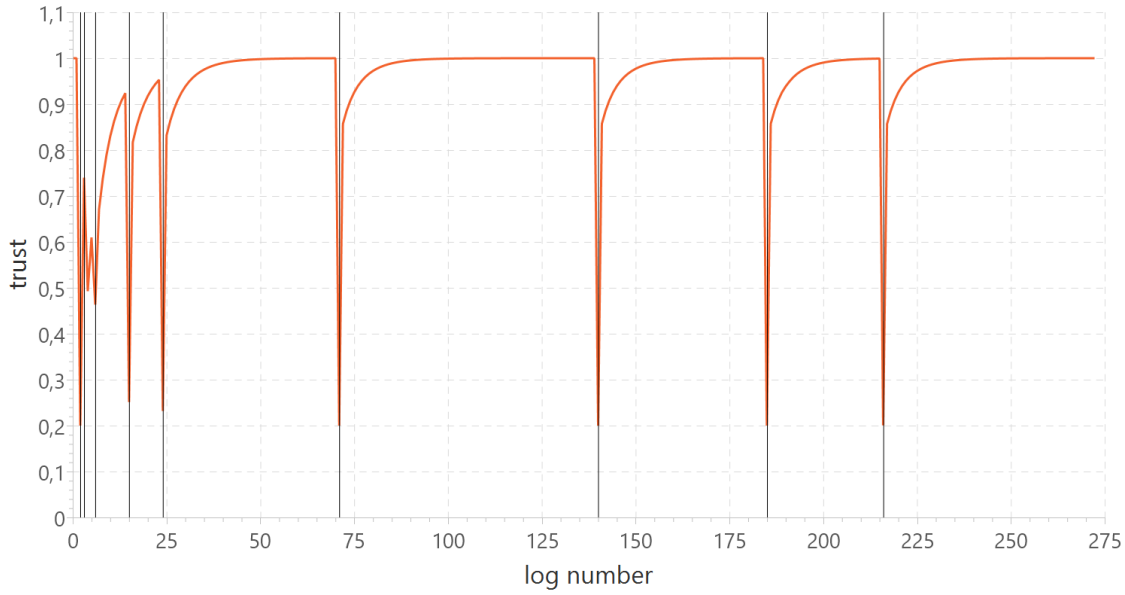
For both periods, reliability and competence (Figures 5.5b and 5.5c respectively) the value does not change and remains constant at its maximum level. This behaviour is expected since the injected messages are not recognised as errors. This is due to the fact that the error messages are predefined and part of the existent messages.

Likelihood is shown in Figure 5.5d. It is a key dimension in decreasing trust for noise in this scenario as in most cases its value decreases significantly to 0.1, e.g. entries 71, 140 or 185. Another situation can be observed when two consecutive messages are noisy (entries 2 and 3): since the first one was successfully recognised as unlikely, its likelihood value, and thus trust value, are low. Now even though the next entry is also noisy, the likelihood is not decreased significantly, as at this point it is unknown whether the lack of necessary transition between these two messages is due to the low trust message or to the next, unknown one. In this situation, likelihood assumes the first option, thus the likelihood decrease is low.

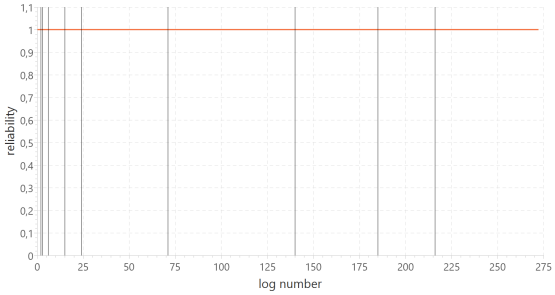
Finally, credibility is illustrated in Figure 5.5e. Its evolution is mostly straightforward. Since the messages do not exist as information normally provided by the considered sensor, there is no confirmation or invalidation specified. Because of that, the credibility value for each encountered noise equals 0.5 as it is neither confirmed or invalidated. In addition we can observe that a low trust for the noisy messages still affects neighbours which then affect the following messages after a noisy log entry: similarly as in the previous scenarios, this behaviour is illustrated by the delayed full recovery for credibility after the problematic message and it is caused by a lack of full confirmation.

5.3.7. Conclusion

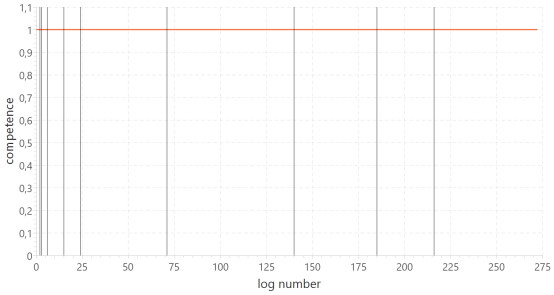
Considering four distinct realistic scenarios of problematic log files, the illustrative results described in this section show that the proposed ReCLiC model has the expected behaviour: it is indeed able to assign low trust values to the simulated noisy entries, it also offers the desired



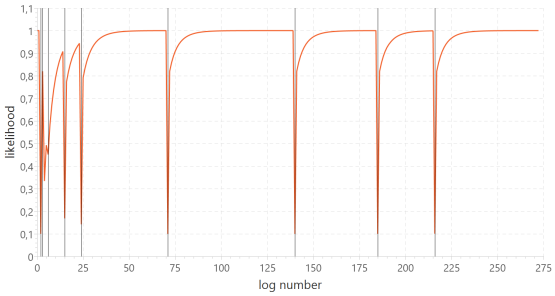
(a) Trust



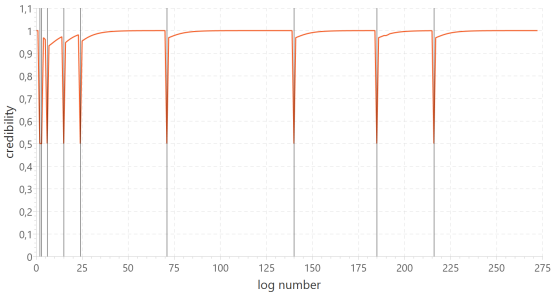
(b) Reliability



(c) Competence



(d) Likelihood



(e) Credibility

Figure 5.5: Non-existent message injection

slow recovery behaviour. Moreover, the experiments show that all four dimensions have a role to play, in different cases, for the global trust model to have this expected behaviour.

Even if they remain a preliminary study to be completed and generalised, the obtained results allow to validate the proposed ReCLiC model and consider its application to real data to detect issues encountered by sensors. Before this study, described in chapter 6, the next section examines in more details the effects of the parameters the ReCLiC model relies on.

5.4. Experimental study of the ReCLiC parameters

As described in Chapter 3, the proposed ReCLiC model mainly depends on three parameters: first the two-step combination of the four dimensions into the trust score relies on the choice of an internal aggregation step that performs a preliminary combination of three dimensions into a single value; the trust score then depends on the weighting parameter to compute the weighted average of this value with the fourth dimension, denoted α. Third, ReCLiC also depends on the time window used to compute reliability and competence, denoted w. These parameters allow ReCLiC to be adapted and personalised to the user's need and the processed data. This section proposes to study experimentally the effect of these three parameters, using the same type of simulated data as the previous experiments, for the same reasons as discussed in Section 5.1, p. 81. This section first describes the considered noise scenario and then studies, in turn, each of the three parameters.

5.4.1. The proposed noise scenario

In Section 5.2, different scenarios for noise simulation have been presented. In Section 5.3 we observed that all dimensions are necessary to score properly trust in various scenarios. Since these scenarios impact trust scoring in different ways, instead of choosing one to analyse the effect of ReCLiC parameters, we propose to combine them. The scenario considered in this section starts with a short burst noise with $\eta = 90\%$ for 15 entries and applies a uniform noise with $\eta = 3\%$, including noise injection of regular messages produced by the considered sensor as well as non-existent ones. By using such a combined scenario, any bias towards one dimension is not possible, which is crucial for the parameters related to the aggregation of the four dimensions into trust, namely the choice of three dimensions that are internally aggregated and the value of the external aggregation weight α.

To study the time window parameter, we consider a different scenario: since the time window only affects reliability and competence, the only important message is *disturbed*. As a consequence for this parameter studied in Section 5.4.4, p. 104 the simulated noise only contains one type of message (*disturbed*) and it is based on uniform noise with $\eta = 3\%$ and an addition of a burst noise episode with $\eta = 90\%$ within 15 log entries. In both cases, as for the study

performed in Section 5.3, except subsection 5.3.2, noise applies to a single sensor and a single topic.

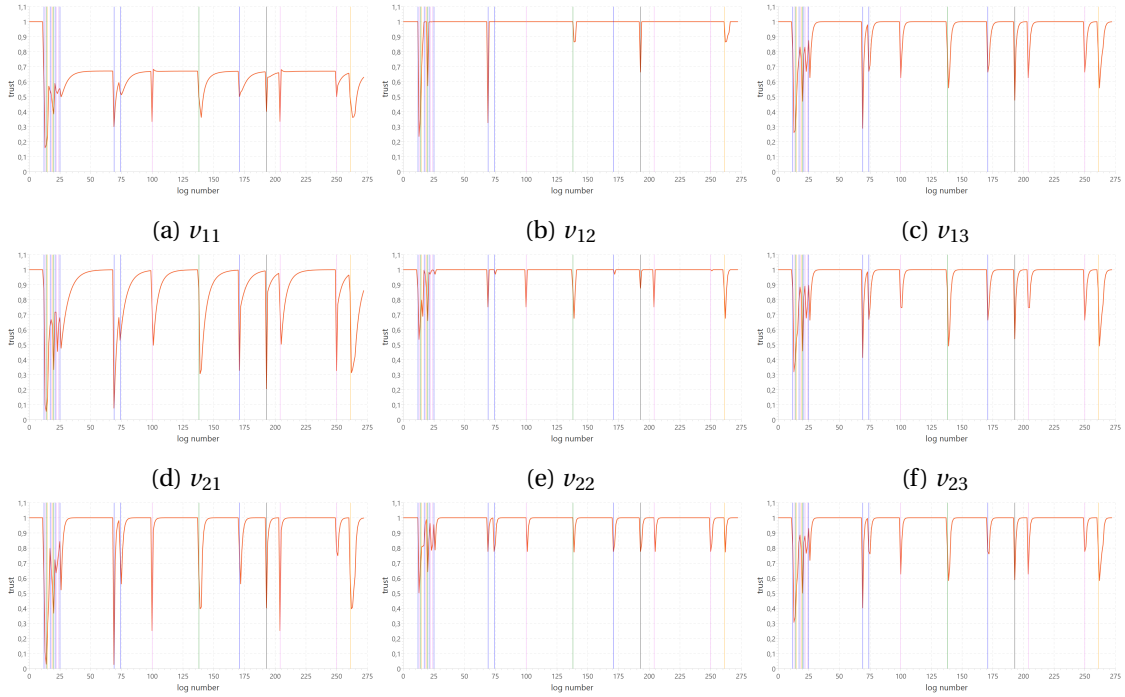
5.4.2. Study of the internal aggregation operator

As discussed in subsection 3.7.1, p. 58, there are multiple possibilities to combine the four considered dimensions into trust. ReCLiC applies a two-step process that performs a preliminary combination of three of the dimensions, whose output is then combined by a weighted average with the fourth dimension. This section focuses on the first preliminary step and studies both the type of operator and the considered three dimensions: it examines all 9 possibilities given in Table 3.1, p. 59, reproduced in Table 5.1, p. 100 to ease interpreting the graph shown above in Figure 5.6. It first analyses the operator type and then the dimension arrangement. For this study, the second step of the aggregation is fixed to the default definition of ReCLiC i.e. considering a weighted average with $\alpha = 0.75$ (subsection 5.4.3 studies the effect of varying α for a fixed first step). Figures 5.6a-5.6i show the results obtained by the 9 aggregation variants for the same simulated data.

Study of the aggregation type The three columns in Figure 5.6 correspond to three different approaches to aggregation, respectively: conjunctive, disjunctive and compromise. We can observe that the general overview of trust decrease and recovery is similar in all three charts of each column, thus general properties for each aggregation type can be highlighted.

In the case of conjunctive behaviour, illustrated in Figures 5.6a, 5.6d and 5.6g, the decrease for noisy entries can be described as the most pessimistic since the trust level has the lowest values compared to the other aggregation types. This is the preferred behaviour since the goal is to highlight all possible problems in the data as clearly as possible. This also includes the dense noise part where the trust values in this area are low during that time, hovering around 0.5. We can observe that in other aggregation types the trust value does not decrease as much for this type of noise. The conjunctive behaviour also shows the biggest differences between the three-dimension arrangement. The first configuration (Figure 5.6a) can be considered as too pessimistic as the maximal trust value is around 0.6 only. This is discussed more in the next paragraph.

The disjunctive behaviour is illustrated in the second column, in Figures 5.6b, 5.6e and 5.6h. We can observe that, of course, the trust values are globally much greater than with the conjunctive aggregation. Moreover, the expected trust decreases for the noisy log entries are small or even do not take place (see e.g. Figure 5.6b, which represents the most optimistic behaviour for log entries 100, 171 or 250). For the configuration shown in Figure 5.6h, except in the initial burst noise period, all noisy entries have the same effect on trust that decreases from 1 to 0.78, whereas the conjunctive aggregation represented in the first column offers a more contrasted behaviour. For these reasons, a disjunctive aggregation operator does not appear



(a) v_{11} (b) v_{12} (c) v_{13}

(d) v_{21} (e) v_{22} (f) v_{23}

(g) v_{31} (h) v_{32} (i) v_{33}

Figure 5.6: Trust evolution for different internal aggregation operators and configurations. $v_{i,j}$ is defined in row i and column j of Table 5.1 below

	Conjunctive (j = 1)	**Disjunctive (j = 2)**	**Compromise (j = 3)**
i = 1	$\alpha \cdot lkl \cdot cr$ + $(1-\alpha) \cdot (r \cdot co)$	$\alpha \cdot (lkl + cr - lkl \cdot cr)$ + $(1-\alpha) \cdot (r \cdot co)$	$\alpha \cdot (\frac{1}{2} lkl + \frac{1}{2} cr)$ + $(1-\alpha) \cdot (r \cdot co)$
i = 2	$\alpha \cdot (r \cdot co) \cdot lkl$ + $(1-\alpha) \cdot cr$	$\alpha \cdot ((r \cdot co) + lkl - (r \cdot co) \cdot lkl)$ + $(1-\alpha) \cdot cr$	$\alpha \cdot (\frac{1}{2}(r \cdot co) + \frac{1}{2} lkl)$ + $(1-\alpha) \cdot cr$
i = 3	$\alpha \cdot (r \cdot co) \cdot cr$ + $(1-\alpha) \cdot lkl$	$\alpha \cdot ((r \cdot co) + cr - (r \cdot co) \cdot cr)$ + $(1-\alpha) \cdot lkl$	$\alpha \cdot (\frac{1}{2}(r \cdot co) + \frac{1}{2} cr)$ + $(1-\alpha) \cdot lkl$

Table 5.1: ReCLiC final aggregation step (copy from Table 3.1, p. 59), implemented with $\alpha = 0.75$ for Figure 5.6

to offer a desirable behaviour for sensor information scoring: it can only highlight major issues that are reported by all the considered dimensions. This behaviour may be useful in case of high noise scenarios, where there is a need to differentiate more severe problems from less important ones.

The case of compromise aggregation is illustrated in the third column, Figures 5.6c, 5.6f and 5.6i. As the name suggests, it offers a compromise between the two previous behaviours. As the conjunctive operator, it recognises all noisy entries; however, the trust values do not decrease as much in these situations. On the other side, as the disjunctive operator, it can highlight more severe problems and differentiate them by decreasing trust stronger. The difference between the three configurations is more subtle compared to other aggregation types and they mainly influence the decreased amplitude for the noisy entries.

Study of the dimension arrangement Different positions of dimensions in trust aggregation can be compared by analysing rows in Figure 5.6: it can be observed that the position of the dimensions affects the level of decrease for each type of noise more than the overall trust evolution. It means that by choosing which dimensions should be combined together we can change the level of trust reactions for different scenarios. For instance, in first column, in Figures 5.6a, 5.6d and 5.6g, entry 100 has trust level 0.34, 0.5 and 0.26 respectively, depending on the dimension arrangement.

In the last column, Figures 5.6c, 5.6f and 5.6i have a very similar trust evolution, including trust decrease and recovery. This results from the compromise operator where all aggregations between dimensions are averages. Even if the arrangement changes, the global aggregation is a weighted average, the only difference comes from the weights. However, these weights can take only values $\frac{\alpha}{2} = 0.375$ or $1-\alpha = 0.25$ with the value $\alpha = 0.75$ considered in this section. The difference between these weight values is not enough to lead to significant differences in the final trust values.

Conclusion This preliminary study supports the choice of defining ReCLiC using a conjunctive operator with the second dimension arrangement, corresponding to the aggregation illustrated in Figure 5.6d. Indeed, the large trust decrease for all noisy entries is desirable, including slower recovery and trust decrease for dense noise. Moreover, the chosen aggregation in the first step evaluates the internal state of the sensor (through reliability, competence and likelihood) and then, in the second step, adds external information provided by other, neighbouring, sensors, through credibility. This study justifies the proposed definition stated in Equation 3.9, p. 59 and recalled here:

$$trust = \alpha \cdot r \cdot co \cdot lkl + (1-\alpha) \cdot cr \tag{5.1}$$

5.4.3. Study of the weighted average parameter

This section studies the second step of the aggregation performed by the ReCLiC model to define the trust score from the four considered dimensions. As discussed in Chapter 3, the second step is a weighted average, applied to the product of reliability, competence and likelihood on one hand, and credibility on the other hand. The α parameter thus balances the influence of the individual sensor dimensions and the dimension that depends on other sensors. The data considered for this study is the same as in the one for the internal aggregation operator, so as to ease the comparison.

In this analysis, three different α values are considered: 0.25, 0.50 and 0.75. The greater the value, the more impact is assigned to the internal aspects and less to the external one. The trust evolution obtained for these three values is plotted in Figures 5.7a, 5.7b and 5.7c. When comparing the three figures we can observe that the overall trust evolution is similar. Differences can be noted when analysing the level of trust decrease for noisy entries as well as the speed of their later recovery.

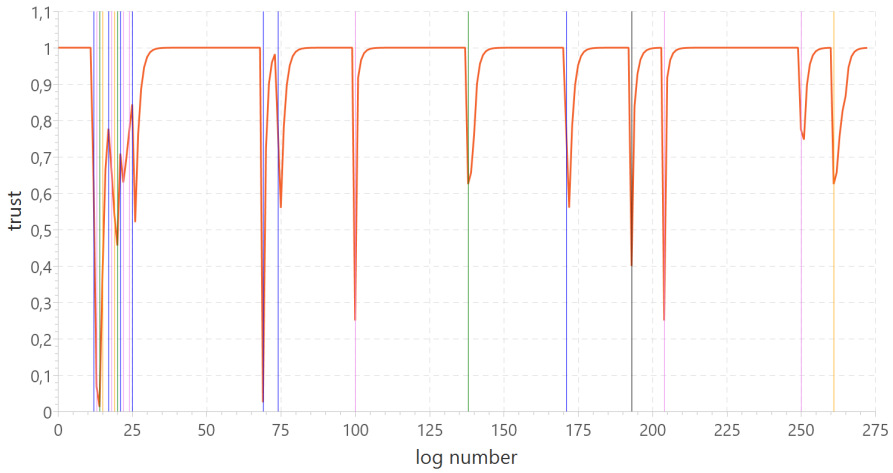
As discussed in Section 5.3, the four dimensions ReCLiC relies on have different behaviours when encountering different noise types. This is reflected in the α value analysis: a simulated problem is recognised differently depending on the α value that gives more influence to some dimensions than others. For instance, the noisy entry 100 has a lower trust value for $\alpha = 0.25$ than $\alpha = 0.75$ but the opposite behaviour is observed when considering entry 250.

More consistency can be observed when analysing error-type messages represented by green and orange lines. For instance, both for entries 134 and 261, the trust value is significantly lower for $\alpha = 0.75$ than for the lower α values. Indeed $\alpha = 0.75$ puts more importance on the internal set of dimensions, that include reliability and competence which depends on error messages. Along the same line, when considering for entry 194, which corresponds to the injection of a non-existent message (black line), the greater decrease can be observed for $\alpha = 0.75$ as the trust value is mostly based on likelihood which is one of the internal dimensions as well.

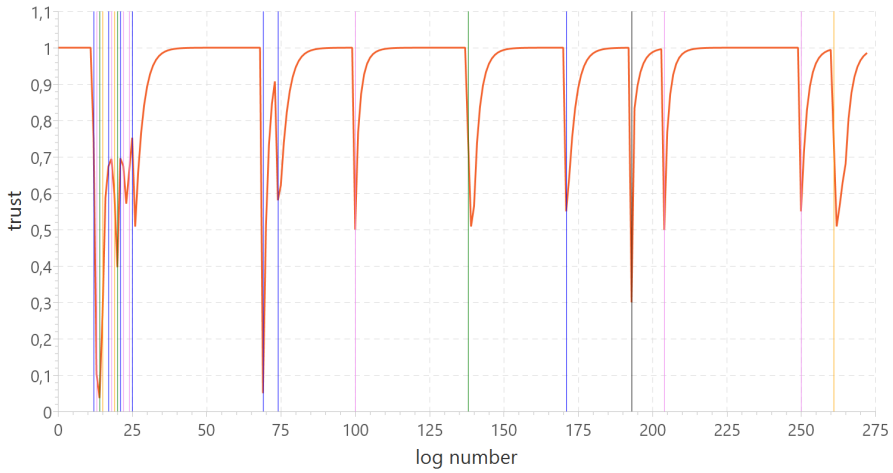
Using $\alpha = 0.50$ provides an interesting behaviour that can be interpreted as an intermediate between the two other options. In this case, numerous noisy entries have an equal trust decrease around 0.5, e.g. entries 100, 139, 171, 204, 250 and 262. These entries vary significantly in trust level in both other approaches.

Another difference is illustrated by the recovery time. The greater the α value, the longer it takes to recover after a decrease. This shows that the recovery time is more correlated with the internal dimensions than with credibility and its duration can be modified by modifying the α value.

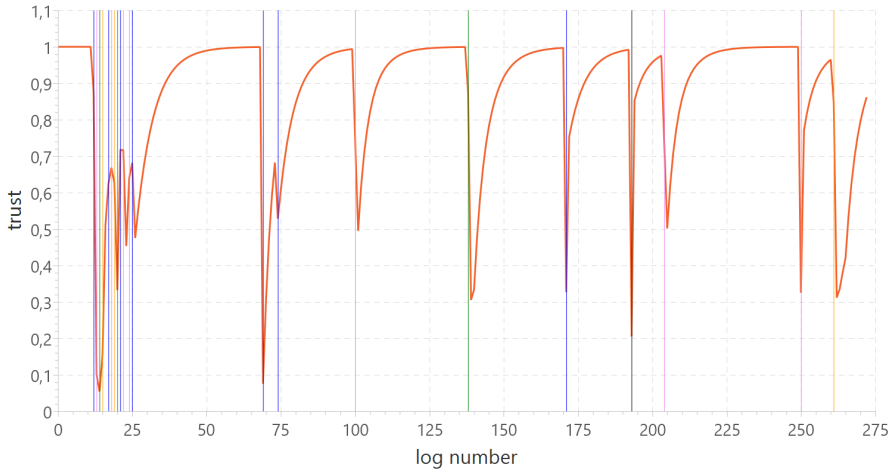
Conclusion In this preliminary study of three α values, the approach taken in ReCLiC is supported. The considered $\alpha = 0.75$ provides more importance for the internal dimensions: reliability competence and likelihood. In most cases, these three dimensions are able to decrease



(a) $\alpha = 0.25$



(b) $\alpha = 0.50$



(c) $\alpha = 0.75$

Figure 5.7: Trust evolution for different α values

trust value in various situations of noisy entries which is a desired feature. The remaining credibility is then able to correct the preliminary evaluation by further decreasing trust or increase this initial scoring.

5.4.4. Study of the time window

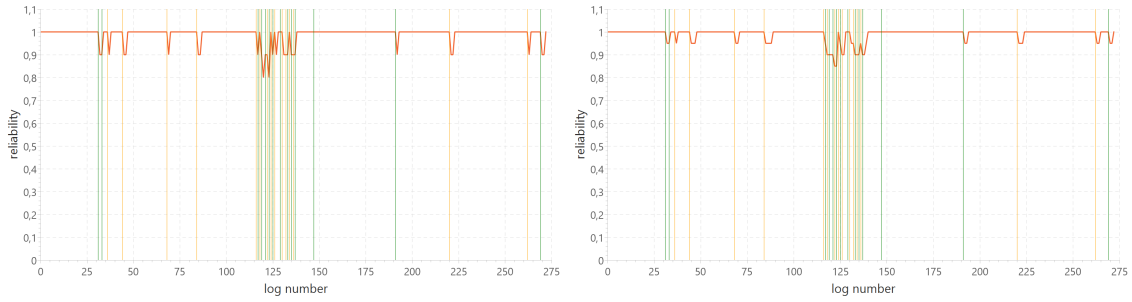
The third ReCLiC parameter we study in this chapter is the size of the time window used to compute reliability and competence (see Sections 3.3, p. 42 and 3.4, p.44 respectively). In this section, four values are compared to observe how they influence reliability, competence and trust: $w = 10$, $w = 20$, $w = 50$ and $w = 100$. They are represented in four charts for each of the three dimensions in Figures 5.8-5.10. In this study, the considered noise is limited to changing the existing entries in the way that its message is *disturbed* instead of *occupied* or *clear*. As in the uniform noise, the amount of changed entries is 3%. In addition, a dense noise scenario is included for entries 116-138 to illustrate how it influences the proposed w values. Entries from a single topic are considered as in the previous studies.

Reliability As discussed and illustrated in Section 5.3.2, p. 86, reliability depends on multiple topics of one sensor. Thus if only one topic produces error values, as considered in this case, the impact on reliability is not significant. However, we can still observe different behaviours when analysing various w values.

Two effects are expected when increasing the time window: the decrease observed for a *disturbed* entry is expected to be smaller, leading to higher reliability values and the duration of the lower value is expected to be longer. Both these effects are indeed observed in Figures 5.8a-5.8d. As discussed in the analytical study in subsection 4.3.1, p. 77, a single decrease or recovery can only take value $\frac{1}{w}$ which gives the biggest single decrease in case of $w = 10$ and the smallest for $w = 100$. However, when considering the duration of the decrease, the bigger the w value, the more entries are subject to a reliability decrease.

Another interesting behaviour can be highlighted for dense noise. We can observe two areas when noise is dense, a large one (entries 116-137) and a small one (entries 31-44). For the latter, for $w = 10$ and $w = 20$, the decreases are independent of each other: the noise is not considered as dense enough. On the contrary, for greater time windows ($w = 50$ and $w = 100$) a combined effect is observed, leading to added decreases. However, even though higher time windows stack reliability values, they are still lower than a single decrease for $w = 10$ or $w = 20$. This can be explained using the analytical study performed in subsection 4.3.1, p. 77: a single decrease for $w = 10$ is equal $\frac{1}{w} = 0.1$. To reach the same level of decrease for $w = 100$, we need 10 consecutive error messages, thus more errors need to be observed to reach a high-reliability decrease.

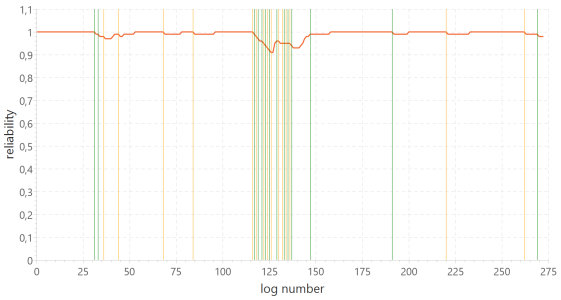
In the case of the longer interval of dense noise (entries 116-137), the situation is similar: stacking decreases from the error entries in larger time windows do not decrease reliability as



(a) $w = 10$

(b) $w = 20$

(c) $w = 50$



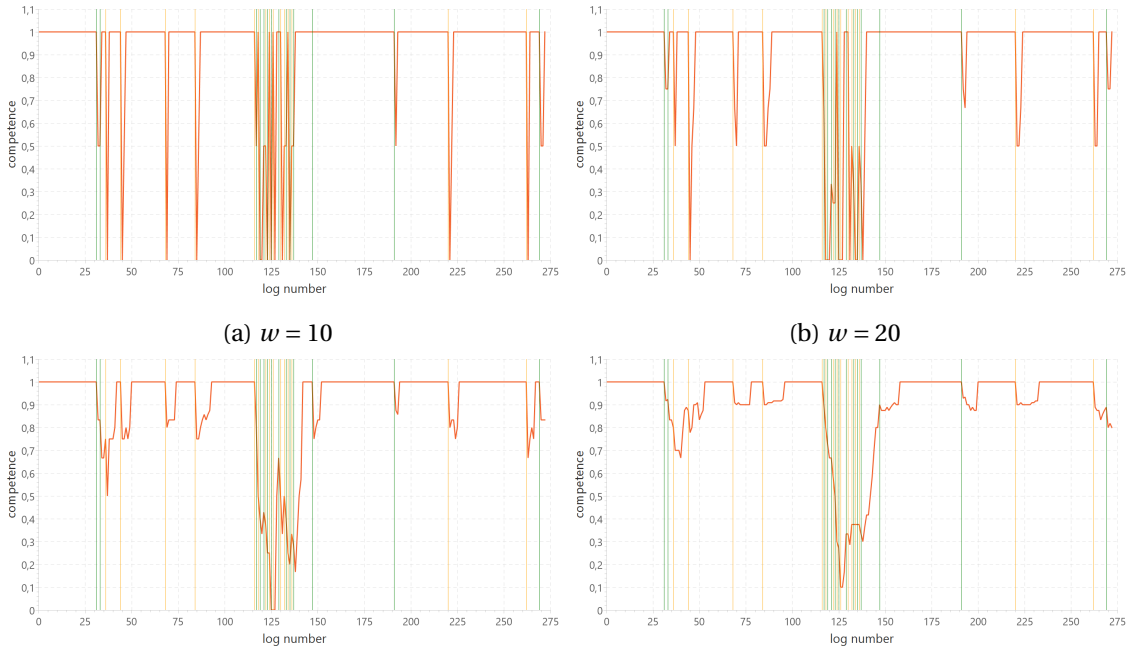
(d) $w = 100$

Figure 5.8: Reliability evolution for different time windows

much as for the lowest window $w = 10$. However, we can observe the difference when analysing the continuity of trust decrease during the entire dense noise period. When the time window is greater, the decrease in reliability is maintained throughout the entire period, whereas for the lower time windows it is more common to decrease quickly and then recover as fast. This allows reliability to get back to the highest value, $r = 1$, within dense noise episode, which does not correspond to the desired behaviour: too small a time window makes the ReCLiC model too dynamic, it does not take into account the history and previous messages enough. A greater value appears to be preferable. On the other hand, too great a value makes the model too inert and not adaptive enough. The appropriate value depends on the dynamics of the system the ReCLiC model must be applied to. For the case of railway application we consider in this thesis, more precisely, for the MoTRicS2015 dataset, we choose $w = 20$.

Competence Competence depends on errors regarding the considered topic, thus the impact of errors is significantly bigger for the competence score than for reliability (see Section 3.4, p. 44). The competence evolution is illustrated in Figures 5.9a-5.9d that show that the size of the time window strongly impacts the possible decrease in competence.

In the case of an isolated error message, the difference is more striking. For the smallest time window ($w = 10$) the competence level often reaches 0, for the largest one ($w = 100$), it hovers around 0.9, e.g. entries 69, 85 or 220. As in the reliability case, the larger time window, the longer the period of competence decrease. It can be observed for entries 69, 85 or 220 where for $w = 10$



(a) $w = 10$ (b) $w = 20$

(c) $w = 50$ (d) $w = 100$

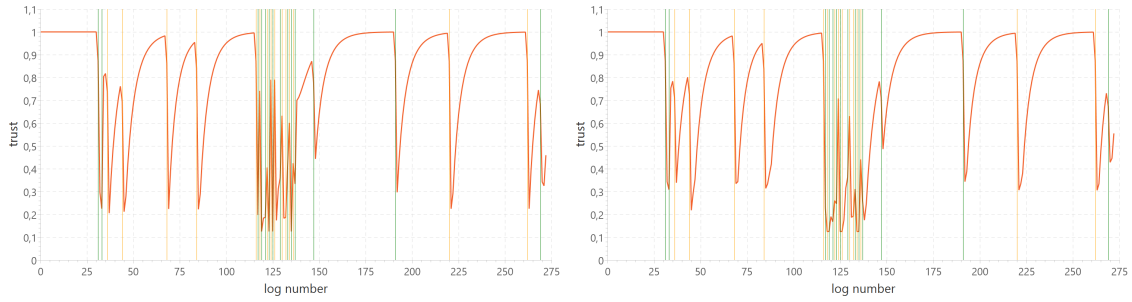
Figure 5.9: Competence evolution for different time windows

competence stays low only for one message and in $w = 100$ stays decreased for around 10 more entries. This window changes accordingly in the intermediate values of $w = 20$ or $w = 50$.

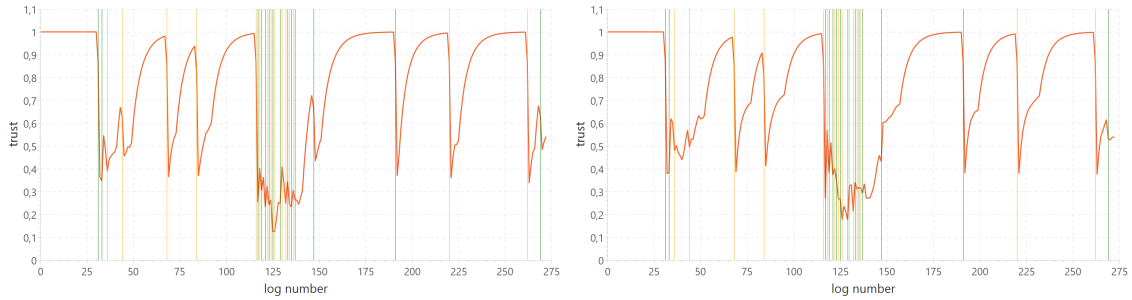
When considering the dense noise period, for all w values, competence is decreased significantly, up to $co = 0$. However, as expected, the decrease fluctuates less in the case of larger time windows as opposed to the lower ones which characterise fast decreases as well as fast recoveries.

Choosing the most relevant time window depends on the expectations. It is necessary to highlight that for higher time windows the decrease of an isolated error is very small. This situation makes the competence value very similar to the reliability one, which is not desired as the two dimensions should score different behaviours. However, it opens the possibility to highlight only dense noise situations. When using lower w values, each error message for the considered topic decreases competence significantly, which is often the preferred behaviour. Thus, as in reliability, the proposed time window for competence is $w = 20$. Having the same w value for reliability and competence is also desirable, as discussed in Section 3.4, p. 44.

Trust The final analysis is performed on trust evolution when modifying the time window, illustrated in Figures 5.10a-5.10d. Interestingly the difference in the level of decrease for the isolated error message is not that significant. When considering entries 69, 85 or 220, the difference in trust between the smallest and the largest time windows is around 0.18. This results from the influence of other dimensions which are part of the ReCLiC model.



(a) $w = 10$ (b) $w = 20$



(c) $w = 50$ (d) $w = 100$

Figure 5.10: Trust evolution for different time windows

A more significant effect can be observed on the recovery time that has a different shape for each considered time window. We can notice that the larger the time window, the more highlighted the two-step recovery: for instance for $w = 100$, after entry 85, trust slowly reaches value 0.72 and then increases the recovery speed again to reach its maximum level. This behaviour is only slightly observed for $w = 10$ and it marks the moment when the error entry is outside the considered time window. For $w = 10$ it is fast, but it takes longer for $w = 100$.

In the case of dense noise, the conclusion is similar to the one from competence as expected. Trust decreases significantly in that situation. The difference lies in the level of fluctuation where lower time windows contain rapid decreases and increases whereas larger ones need more error messages to reach the same level of decrease but also recover more slowly.

Conclusion The preliminary study supports the choice of defining ReCLiC using $w = 20$ for both reliability and competence. At this level, the considered window is high enough to mitigate large dynamics of sudden decrease and recovery for both dimensions and trust, which is not desirable as it does not take into account the history and previous entries enough. Too great a value makes the model too inert and not adaptive enough, leading to prolonged decreases of non-simulated entries. The appropriate value depends on the dynamics of the system the ReCLiC model is applied to. For the case of railway application in this thesis, the value $w = 20$ is chosen.

5.5. Conclusions

In this chapter, the experimental study of the ReCLiC model was performed, using realistic simulated data. Four distinct realistic scenarios were proposed to validate the proposed model in various situations. The illustrative results described in this section show that the proposed ReCLiC model has the expected behaviour: it is indeed able to assign low trust values to the simulated noisy entries, it also offers the desired slow recovery behaviour.

They also allow to validate the proposed parameter values for the application to the axle counter in the railway signalling domain and therefore justify its use for real data with unknown potential quality issues. This study is performed in the next chapter.

6. Application to real data: trust dynamics of multiple sensors in the railway domain

The experimental study performed in the previous chapter on realistic simulated data allows to validate the proposed ReCLiC as an information scoring tool for sensor data and in particular its implementation in the railway signalling domain for the axle counter sensor. This chapter presents its application to real data, more precisely to the MoTRicS2015 data set provided by the industrial partner Thales Polska, described in Chapter 4, on page 63.

This chapter focuses on the final trust values, not decomposing trust into the components it relies on: they correspond to the information a user is interested in, to analyse whether the sensors provide relevant information or whether they contain quality issues. In order to do so, a new visualisation is proposed to present trust scores from multiple devices at the same time, as discussed in Section 6.1. The output of the ReCLiC method is presented and the results discussed in Section 6.2.

6.1. Proposed visualisation of the trust evolution

In the previous chapter, the analyses were performed on a single section of a single sensor. This approach is ideal for a thorough analysis of how trust behaves in a variety of situations of interest. The case of multiple sections and multiple sensors over a long time that occurs in real case scenarios raises several challenges, as discussed in the next subsection. The general principles of the visualisation we propose is presented in Section 6.1.2 and the specific question of time aggregation is discussed in Section 6.1.3.

6.1.1. Challenges of real data visualisation

The visualisation used throughout the experimental study presented in the previous chapter makes it possible to analyse in details the behaviour of the proposed model, but it limits the number of messages that can be presented on a graph. More precisely, only 275 messages related to a single section can be represented, which is very low in comparison to the total number which is around 25,000 messages per year: 275 messages only correspond to four days of activity.

The second issue comes from the total number of sections that are present in the data. The current visualisation focuses only on one section of one sensor, but in this study, a larger scale needs to be considered, as there are 57 sections covered by 5 sensors. This means that 57 charts need to be plotted in order to observe the trust evolution of all sections and their corresponding log entries (see subsection 4.1.2, p. 63 for the presentation of the real data).

In addition, charts based on the entry number instead of time make it difficult to observe how trust behaves among neighbouring sections at the same time. This is due to different time delays between two log entries that can be seconds or hours. The visualisation used in the previous chapter allows to study trust evolutions very precisely as it simplifies the x-axis by using the entry number instead of the exact time (as discussed in subsection 5.3.1, p. 85). However, this approach needs to be changed when considering several sections and sensors at the same time, because of the need to align them in time to observe how a message about one section can influence a message about another.

6.1.2. Proposed heatmap visualisation

To address these challenges, a heatmap visualisation is proposed. In its general definition, a heatmap is a graphical representation of data where the individual values contained in a matrix are represented as coloured squares. For the information scoring case, the principle is to build a matrix where each cell corresponds to a time interval, i.e. a discrete date, and a section. The cell then contains an aggregation of the trust scores associated with the log entries concerning this section in the considered time interval.

In other words, in the heatmap, the x-axis represents the message date and time, the y-axis represents the section of the message and a trust score of the message is represented by a colour between yellow, corresponding to $trust = 1$, and red, corresponding to $trust = 0$. If there is no message produced for a section in a specified time, the absence of trust scores is represented as white space. Choosing the red colour for the low trust messages is relevant for security reasons: red immediately draws the attention of the user, which is important to deal with a potential problem that can cause a significant train traffic turbulence or accident.

6.1.3. Trust temporal aggregation

Using time in the x-axis instead of the entry number is challenging as there is no easy way to propose a valid time interval (see subsection 5.3.1, p. 85). In order to show all messages and their trust scores individually, the time interval would have to be one second: indeed, when a train passes one section, it triggers two messages (*occupied* and *clear*) within a very short time window. However, most of the time, a train passes only once or twice per hour, thus the remaining values would be non-existent and coloured white. In addition, having such a low interval considerably limits the maximum possible length of the chart to a few days.

In order to solve this issue, we propose to aggregate the trust scores within an a priori specified time window t. A higher t allows to display a bigger time frame, however a lower t displays a more detailed view.

In MoTRicS2015, sensors produce a low number of entries per hour which makes $t = 1h$ a good compromise between displaying a large time frame and important details. Within one hour there are most often one or two train passages as the considered station has low traffic.

In order to aggregate the trust scores of all log entries associated with the considered section in the considered time window, we propose to apply a conjunctive aggregation operator. Indeed we want to strongly highlight the decrease in trust value within the data as, in this domain, the security of train passages is very crucial. The currently proposed function is the minimum, which prioritises the lowest trust value in the considered time window. By highlighting where the large decrease in trust value takes place, it is then easier to find the problematic messages and analyse the issue.

6.1.4. Visualisation procedure

The procedure for visualising trust values in the heatmap is thus as follows: entries produced about the same section are grouped together. For each group, trust values within the same hour are aggregated to create a pair *(d,tr)*, where *d* is a date and hour for this aggregation and *tr* is an aggregated trust value. A heatmap representation of the resulting matrix is then applied: each trust value is translated to colour between red and yellow. Then, for each group, their pairs are plotted in the heatmap.

6.1.5. Section order

The order of the displayed sections is another issue. By defining credibility as confirmation from neighbouring sections, the trust scores associated with messages about one section impact trust of the other ones. Moreover, reliability is scored in a way that an error from one section can impact other sections of the same source as well. This means that in particular errors can have an impact among multiple sections of the same sensor which are close to each other. As a consequence, the order of the sections in the heatmap matters.

One possibility to define the section order in the heatmap is to choose the first section and use its neighbour as the next one. This gives the opportunity to observe behaviour which affects multiple neighbouring sensors. All 57 sections can be divided into three groups. The first two represent sections in the same set of tracks, for trains going in each direction, the third set contains the rest, which mostly corresponds to backup tracks. Most traffic can be observed in the first two groups that are responsible for 39 sections.

An example of a heatmap in presented in Figure 6.1. Each hour on x-axis is one pixel wide and its colour represents the aggregated trust score of the messages produced in this period

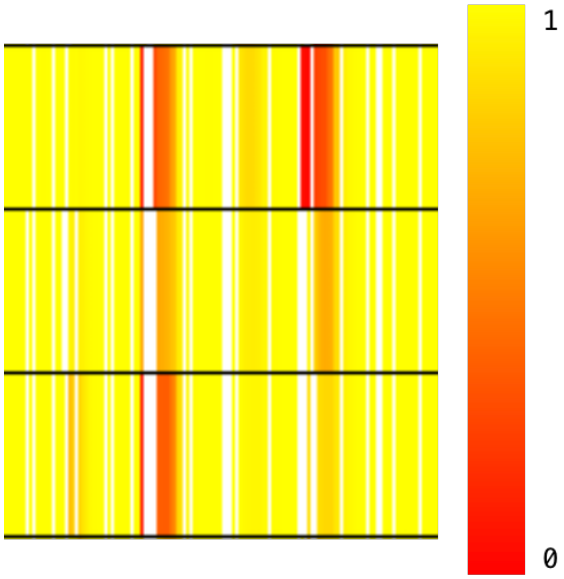


Figure 6.1: Example of a heatmap which contains trust values about three different sections. Red represents the lowest trust score and yellow the highest one. White space represents the absence of messages.

about that section. Sections on the y-axis are separated by a black line where neighbours in the graph correspond to neighbours in the network of sensors. A thicker black line is used to separate two sections that are not neighbours, differentiating the three main groups mentioned above.

6.2. Result analysis

The ReCLiC model is applied to the MoTRicS2015 database with the default parameters (see Chapter 5): the four considered dimensions are scored as recapped in Table 4.2, p. 76 with $\alpha = 0.75$ and $w = 20$. The considered state transition graph is provided by the expert, as shown in Figure 4.4, p. 69 and the results are visualised using the heatmap as discussed in the previous section.

The considered time period is 3000 hours, between March 1st and July 4th 2015. This period is much larger than the one considered for simulated data in Chapter 5, which was limited to a few days. In addition, the period from Chapter 5 was chosen so that no quality issue occurs, so as to control their introduction with noise simulation. In this chapter, the goal is to find potential issues in the data, thus no such limitation is considered.

In the following subsections, the outcome of the ReCLiC application to the MoTRicS2015 database is studied. First, a general view of the heatmap is discussed in subsection 6.2.1. Then, the low trust areas are discussed in subsection 6.2.2. Finally, the effects of low trust propagation on multiple sections are studied in subsection 6.2.3.

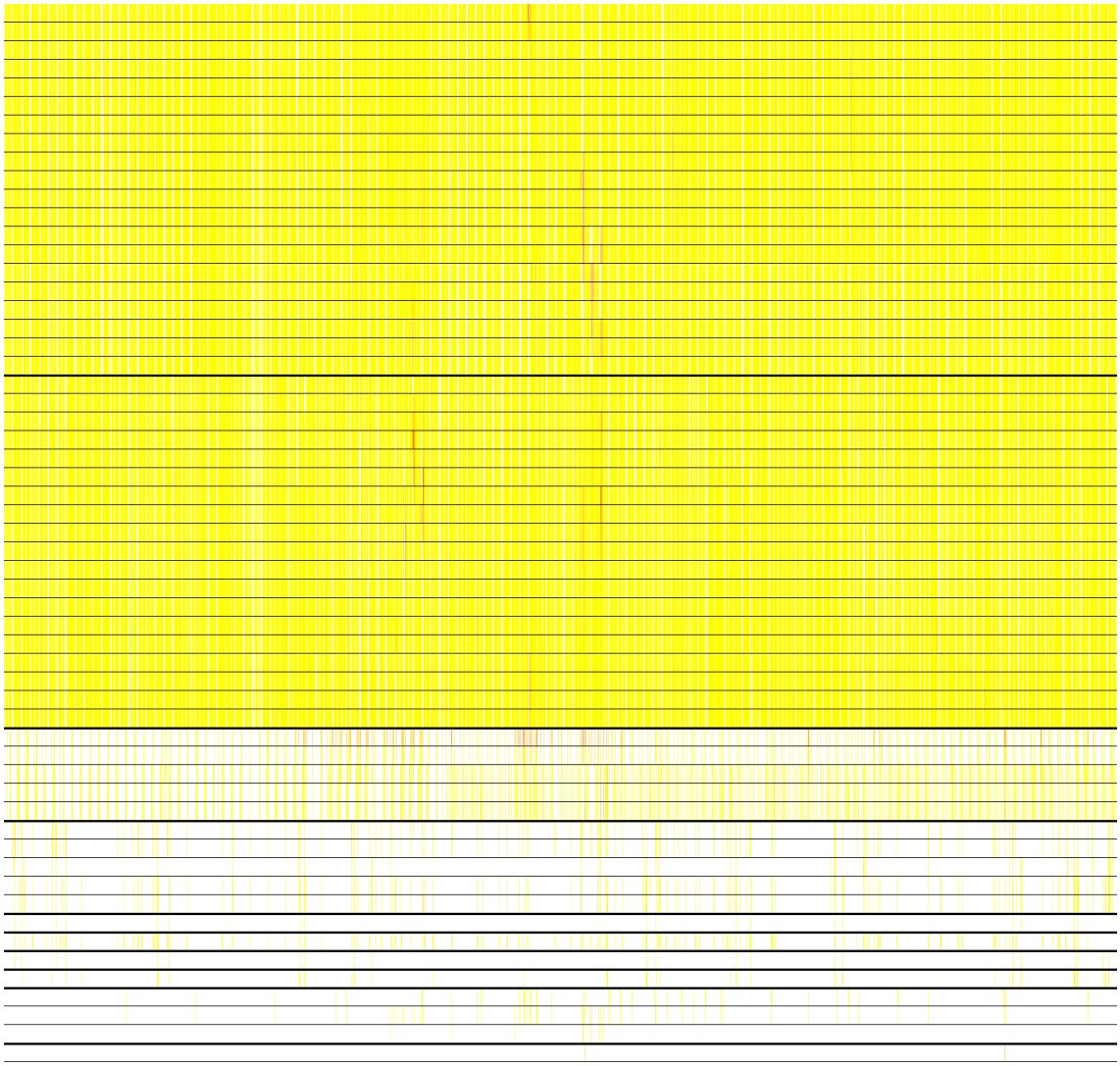


Figure 6.2: Trust value evolution for 57 sections of the real dataset MoTRicS2015.

6.2.1. Global view

Figure 6.2 shows the obtained results. We can first observe the third group of sections in the lower part of the graph that corresponds to backup tracks: they are much less active than the others, which implies that white is a dominant colour for them. However, they do not all show the same behaviour, some of them are more active than others and their trust evolution varies as well. For instance, the first five sections in this group are more active than the others, in addition, the first one has a low trust score for most messages, unlike the others. In opposite, we can also see sections that have almost zero traffic for the entire considered time interval.

The two other groups are separated by a thicker black line. Each of them corresponds to a sequence of sections which are triggered by a train travelling on main tracks. The two groups respectively relate to the two directions a train can travel. Because of that, all sections from one group are neighbours and we can observe how a low trust message in one section can impact

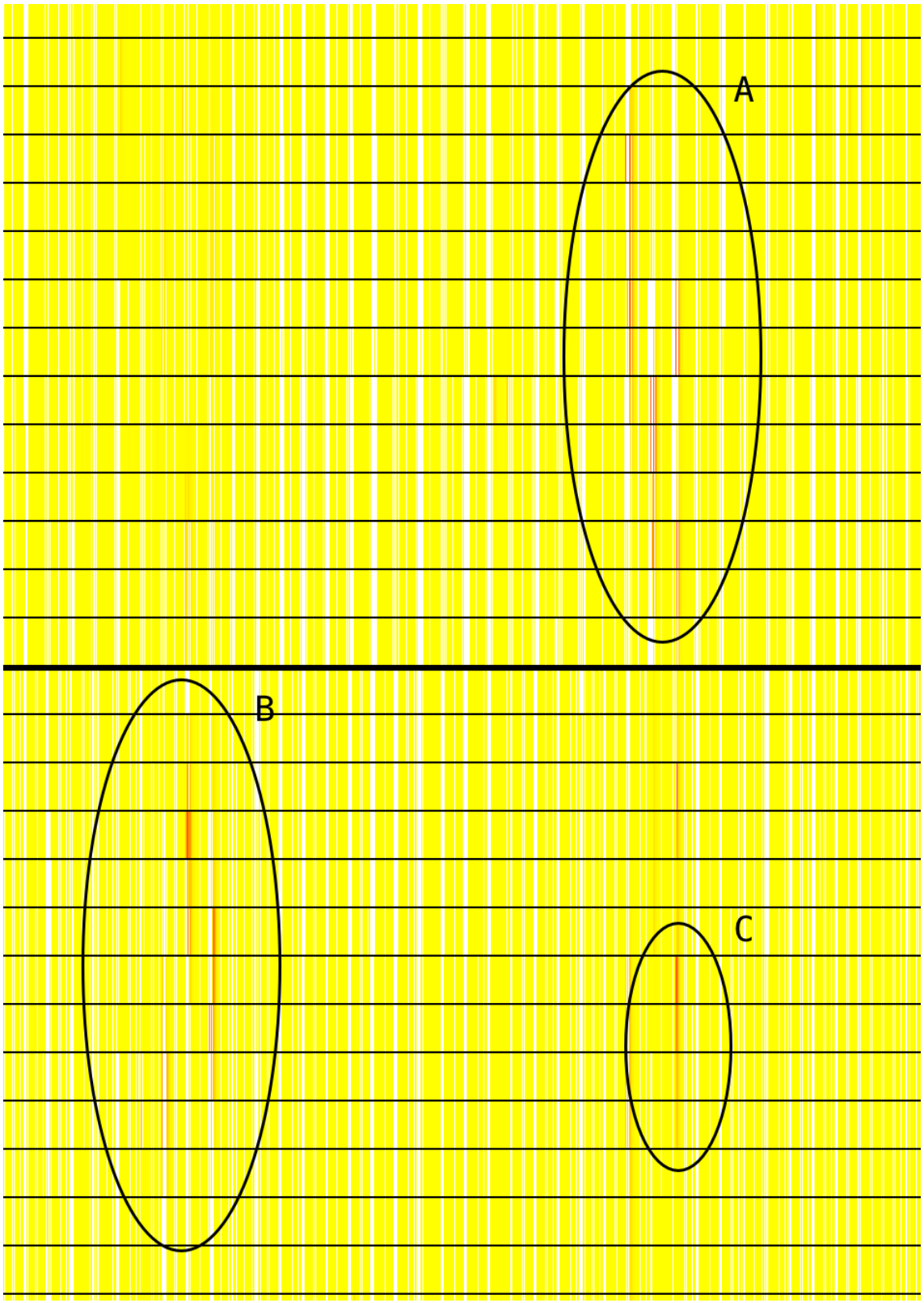


Figure 6.3: Zoom on significant trust decreases over multiple sections.

trust for messages in the others.

6.2.2. Global quality issues

In Figure 6.2 we can observe areas where trust significantly decreases, mostly in the centre of the graph. This part of the graph is extracted and enlarged in Figure 6.3 to better see the details: three important parts are highlighted. The first two, marked as A and B, show various trust decreases that affect multiple sections in a short period of time, the third one, C, is discussed in the next subsection.

The problems which affect quality in MoTRicS2015 decrease trust between two and six sections at a similar time period. Moreover, the trust decreases we observe are not limited to a single message, we can notice that for each case, messages within multiple hours are affected by low trust, highlighted by the red colour. This means that the problem that caused this major decrease in trust was not resolved immediately.

In addition, we can see that the sections do not provide messages every hour and there are white spaces frequently, however, not in the amount that can be observed in the backup sections. Noticeably, in many cases, white cells are present for the same hour in multiple sections. In addition, they can be found at similar intervals which might allow to recognise between day, where trains travel normally, and night where the train traffic often stops completely.

6.2.3. Quality issue propagation effects

Part C in Figure 6.3 illustrates a different case: even though it also presents a trust decrease affecting several sections, the covered time period is much smaller. To get a closer look, this part is extracted and enlarged in Figure 6.4 where a part of six sections is illustrated.

We can observe that in the second section a major trust decreases is present for multiple messages that are produced in several hours interval. It is illustrated by the intensive red colour over multiple consecutive cells in this section. This sudden decrease in one section impacts its neighbours. We can see that the trust values in the same time period for the third section are also reduced, however not as low as for the second section. This propagation continues for the following sections where trust is decreased less each time it propagates to the next neighbour until it is no longer noticeable.

The reason for this behaviour is that the initial trust decrease disrupts a chain of confirmations or creates an invalidation. The credibility component of the ReCLiC method adds an external factor, where messages from neighbours can influence the final trust value by confirming or invalidating the considered information. This shows that a decrease in trust for one section can decrease the trust values of messages from neighbouring sections. Therefore, the neighbour either lacks the expected confirming message or has to consider a low trust one.

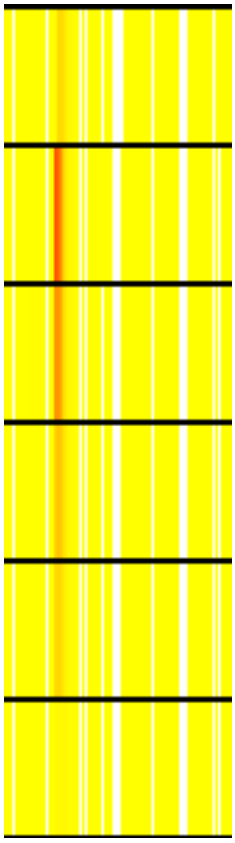


Figure 6.4: Zoom on C from Figure 6.3.

In either case, it leads to a decrease in its trust value and propagates this decrease to another neighbour.

The level of decrease is lower each time it propagates to another section. This level is controlled by the a priori set α value (see Section 4.2.5, p. 75). By manipulating this constant, it is possible to change the level of propagation to the neighbours, which can also affect the number of sections that are affected by the original trust decrease (see Chapter 6).

An interesting fact is that the propagation can be observed stronger in one direction than the other. When considering sections below the one with the original quality problem, the trust decrease is stronger and lasts for three or four following sections. However, the decrease in trust above the considered section is not symmetrical, the decrease is present however in much less intensity.

6.3. Summary

This chapter focuses on a real application of the ReCLiC method in the case of messages produced by axle counters in MoTRicS2015 database. As this dataset was not preprocessed by an expert, its quality is unknown beforehand. In order to perform this analysis, a new visualisation tool was required, so as to visualise information provided by all sensors on all topics at

the same time.

To solve the considered problems, a heatmap is proposed that allows to examine trust scores depending on the section and time. It makes it possible to observe that indeed the proposed information scoring model is able to highlight problematic messages. These trust decreases are observed in small groups which include multiple sections at the same time, suggesting a global problem at that time.

In addition, it can be observed how a decrease in trust for one section can impact messages from neighbouring sections. This impact is not only limited to a single following neighbour but propagates further making a visible chain of low trust messages that recovers with following sections.

7. Conclusion and future works

7.1. Considered problem and challenges

Information quality proves to be very important when considering any kind of information. It allows for a decision-aid system to differentiate between useless and worthwhile information which significantly impacts its capabilities. Information quality assessment highly depends on the type of information, its context and considered domain. It is usually decomposed into different criteria, or dimensions, that allow to capture, measure and combine different aspects of a piece of information.

When the considered information to score is provided by sensors, dedicated information scoring models have been proposed, exploiting the specific properties of such sources. They mainly focus on two criteria, reliability and credibility, although their definition and implementation differ significantly across the models.

Regarding reliability, the main approach is based on the sensor accuracy: it consists in comparing its output with some ground truth to assess whether the pieces of information it provided were correct. The more accurate the sensor was in the past, the higher its reliability score. The problem with this approach is its reliance on ground truth which is often unavailable, expensive or difficult to acquire. In other approaches, assessing a piece of information relies on available meta-information, understood as any additional information to the one provided by a device, e.g. detailing its specification or context. This type of information is often specific to the considered sensor and its usage, and thus difficult to generalise. A method that considers such meta-information is often limited to be only used in the corresponding specific case.

7.2. Contributions

In order to deal with these challenges, this thesis proposes the ReCLiC model to score information provided by sensors and designed to be sensor-generic, not dependent of ground truth and dependent only on easy-to-access meta-information, exploiting only attributes shared among a majority of sensors.

Generic trust scoring framework for sensors: the ReCLiC model Informally, the ReCLiC model takes as input a log file of sensor entries and aims at attaching each log entry with a numerical evaluation of its quality: this quality is understood as the trust that can be put in the message content of the log entry, measured by considering the source, the content and the context of this log entry, which are the three main components defining a piece of information.

The proposed ReCLiC model is named after the four components it integrates, namely Reliability, Competence, Likelihood and Credibility: ReCLiC contains a general criterion, which refers to the source, as well as more specific ones that consider the piece of information itself, as well as its context.

Reliability is the general assessment of the source. By considering recent error messages produced by this source, it assesses its ability to provide any meaningful information in the considered time window. *Competence* also describes the source but more specifically, in relation to the topic of the considered piece of information. Similarly to reliability, it considers recent error messages, however, it examines which topics are associated with these error messages and limits the competence decrease to these topics only. *Likelihood* evaluates the content of the information. It is based on temporal confirmation and it checks whether two consecutive messages are possible according to a state transition graph that can be learnt from the data. *Credibility* is the fourth criterion which looks for confirmation or invalidation of the considered piece of information. It is based on a network of sensors which represents the expected correlation between the available sources, e.g. based on their geographical location.

The four considered dimensions are discussed in-depth and their role in evaluating quality is examined in details. Each dimension and its requirements are commented, and motivated definitions are proposed after considering several possible variants. All dimensions are scored in a way that ensures a generic behaviour and allows to be adapted to any information provided by sensors.

Trust is the result of combining the four proposed dimensions. Several aggregations are considered that differ in the applied operator as well as the position of the dimensions. Their respective semantics are discussed, highlighting the flexibility of the ReCLiC model that can be adapted to different needs of the user.

ReCLiC implementation for axle counters An operational instantiation of the generic ReCLiC definition to a real case is proposed for a specific sensor in the railway signalling domain: the form of the four dimensions for this case is discussed and a formal study of the information score behaviour is performed, analysing each dimension separately.

A real dataset provided by Thales Polska is studied. Methods to extract from the data the two types of required meta-information: the state transition graph and the network of sensors are proposed. This further limits any dependence on external knowledge for the implementation of the generic ReCLiC model.

ReCLiC experimental validation To evaluate the ReCLiC model an experimental study is conducted, based on realistic data simulation which modifies the original dataset to introduce different kinds of noise. The original dataset is then considered as ground truth to assess the ability of ReCLiC to assign low trust scores to the simulated noisy entries. Such a simulation allows to control the introduced types of problems, their quantity and distribution. Moreover, it allows to simulate rare situations which do not occur in the original dataset. Four different scenarios are proposed to simulate various realistic situations.

This experimental study includes a study of the ReCLiC parameters and validates the proposed values, that make it possible for ReCLiC to offer the desired behaviour, in particular with respect to trust decreases and later recoveries.

Real data and trust propagation analysis Finally, the ReCLiC method is used to evaluate real information from the real dataset provided by Thales Polska. The obtained results make it possible to highlight several low trust pieces of information which are analysed and discussed.

In addition, the way how a decrease in trust for one piece of information influences other messages from the same or other sources is analysed, in particular observing how trust propagates to the neighbour sources of an initial source affected by a trust decrease. To do so, a new visualisation method is proposed to show multiple trust scores from many sensors at the same time, allowing to take a dynamic point of view and to perform the analysis of trust propagation.

7.3. Future works

This section discusses directions for future works opened by this thesis. It starts with short term work, based on the extension of the ideas presented in this thesis, and goes to long term approaches of usage and improvement of the ReCLiC model. They include experimental as well as theoretical perspectives.

Noise scenarios and validation criteria Evaluating any trust or quality scoring model is a challenge as it is difficult to provide the expected quality scores to problematic information. Moreover, the studied models often work differently when encountering different types of information in various contexts. The approach presented in Chapter 5 is the proposed solution to this problem: using data simulation in various scenarios allows to validate any trust scoring model, using the original database as a reference.

Future works may include additional scenarios, in particular so as to study the notion of critical threshold: this notion can be understood as the determination of a noise level such that the trust model somehow breaks down and does not manage to recover from low trust values, even when it should do so. This could open the way of discussing and studying a possible notion of the robustness of information scoring models.

Another perspective regards the definition of a numerical validity criterion, to go beyond the qualitative comments of the observed behaviour of a scoring model. This issue is a challenging one: for instance, a basic approach comparing the occurrences of observed trust decreases with that of the noisy entries raises two problems. On the first hand, it may be difficult to define the values that are low enough to be accepted as trust decreases. On the other hand, such an approach does not capture the desired behaviour of slow recovery, which, on the contrary, may be interpreted as incorrect identification of noisy entries and thus as a drawback. A more complex approach that naturally derives from these observations may be to define expected trust values and compute a mean square difference with the obtained values. However, as discussed in the thesis, the definition of such a ground truth is a very difficult task. This makes the proposition of a numerical validity criterion a challenge, although it may be a useful complement to the detailed qualitative analysis performed in the experimental studies.

Expert validation of the real data results The natural continuation of the current work is to analyse the low trust areas of the ReCLiC implementation to MoTRicS2015 database, presented in Chapter 6. We observed areas where the trust score decreases significantly for one or more sections. The exact messages and their context should be extracted and discussed with an expert who could evaluate whether the considered messages are indeed problematic and what was the origin of the problem.

From an operational point of view, a perspective could be the development of an interface and an exploration tool to connect the heatmap visualisation with details about the corresponding log entries and their context, so as to help the expert to investigate the potential issues highlighted by ReCLiC and the heatmap visualisation. This tool could have two main applications: on the one hand, real-time quality control, allowing the human agent to monitor the current state of sensors and quickly act when a problem rises, on the other hand, a posteriori data quality control, allowing to maintain high-quality databases for forensic purposes by reporting and highlighting the low-quality messages to be investigated.

Legibility and subjectivity For the human operator facing the system, one of the challenges in trust scoring, or generally quality scoring, is understanding the obtained results. Among others, the exact level of high/low trust values, the difference between two trust values or the level of potential risk with various trust values may be difficult to assess. This issue is related to the choice of using numerical scores in the interval $[0,1]$, which, on the other hand, provides high flexibility for the score definition. One perspective is to extend the formal study of the respective dimension impact on trust, so as to define strategic trust levels that can be associated with critical responses, differentiating between high and low-quality problems.

Another interesting direction for future work related to user-centred issues regards the integration of the human operator subjectivity regarding the trust definition and the possibility to

adapt the information scoring model in a personalised way. ReCLiC is, by design, a highly flexible information scoring model, where in particular the aggregation operators can be adapted according to the user needs. Beyond the explicit parameters, studied in the experiments on simulated data, it may, for instance, be interesting to study the definition of other aggregations of confirmation and invalidation in the credibility definition: it could for example allow for more cautious users, who could be prone to giving more weight to invalidation than to confirmation. Along the same lines, the competence is defined as taking the maximal value, i.e. $co = 1$ in the case where nothing is known about the considered sensor and topic: other behaviours, in particular, cautious ones, may be investigated.

Multiple types of sensors in different domains In this thesis, the proposed generic ReCLiC model is implemented for the axle counter sensor. Possible continuation of the work includes other sensors in this domain e.g. traffic lights or point machines in a general framework of heterogeneous information fusion. Indeed, the ReCLiC model allows the cooperation of multiple sensors of different types in order to analyse trust for the information they produce. This cooperation is in particular based on credibility where the level of confirmation or invalidation from other sources allows to improve the internal score. A future direction of research is to investigate whether including more types of sources can lead to an improvement in assessing trust levels when encountering low-quality information. Moreover, to further highlight the generic feature of ReCLiC, other domains could be considered to implement the proposed model, showing how to implement each dimension as well as interpret the obtained results.

Bibliography

Appriou, A. (1998). Uncertain data aggregation in classification and tracking processes. In *Aggregation and fusion of imperfect information*, 231–260. Springer.

Appriou, A. (2001). Situation assessment based on spatially ambiguous multisensor measurements. *Int. Journal of Intelligent Systems, 16*, 1135–1166.

Balduzzi, M., Pasta, A., & Wilhoit, K. (2014). A security evaluation of ais automated identification system. *Proc. of the 30th annual computer security applications conference* (pp. 436–445).

Baqa, H., Truong, N. B., Crespi, N., Lee, G. M., & Le Gall, F. (2018). Quality of information as an indicator of trust in the internet of things. *17th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications and 12th IEEE Int. Conf. on Big Data Science and Engineering, TrustCom/BigDataSE* (pp. 204–211).

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer International Publishing.

Besombes, J., & Revault d'Allonnes, A. (2008). An extension of STANAG2022 for information scoring. *Proc. of the Int. Conf. on Information Fusion, FUSION'08* (pp. 1–7).

Bisdikian, C., Kaplan, L. M., Srivastava, M. B., Thornley, D. J., Verma, D., & Young, R. I. (2009). Building principles for a quality of information specification for sensor information. *12th Int. Conf. on Information Fusion, FUSION'09* (pp. 1370–1377).

Blasch, E. (2008). Derivation of a reliability metric for fused data decision making. *IEEE National Aerospace and Electronics Conference* (pp. 273–280).

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *Int. Journal of Intelligent Systems, 18*, 51–74.

Chen, Y., Yu, J., He, A., & Tang, Z.-a. (2016). A method for selecting optimal number of sensors to improve the credibility. *Journal of Sensors, 2016*.

Cvrcek, D. (2004). Dynamics of reputation. *9th Nordic Workshop on Secure IT-systems (Nordsec'04)* (pp. 1–14).

Demolombe, R. (2004). Reasoning about trust: A formal logical framework. *Trust Management*, 291–303.

Destercke, S., Buche, P., & Charnomordic, B. (2013). Evaluating data reliability: An evidential answer with application to a web-enabled data warehouse. *IEEE Trans. on Knowledge and Data Engineering, 25*, 92–105.

Detyniecki, M. (2001). *Fundamentals on aggregation operators*. Doctoral dissertation, Technical report, University of California Berkeley and University Pierre and Marie Curie.

Dubois, D., & Prade, H. (1988). *Theory of possibility an approach to computerized processing of uncertainty*. Plenum Press, New York.

Duma, C., Shahmehri, N., & Caronni, G. (2005). Dynamic trust metrics for peer-to-peer systems. *Proc. of the 16th Int. Workshop on Database and Expert Systems Applications* (pp. 776–781).

Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. *Proc. of the 3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2004)* (pp. 740–747).

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management, 39*, 109 – 116.

Florea, M. C., & Bossé, É. (2009). Dempster-Shafer Theory: combination of information using contextual knowledge. *Int. Conf. on Information Fusion, FUSION'09* (pp. 522–528).

Florea, M. C., Jousselme, A.-L., & Bossé, É. (2010). Dynamic estimation of evidence discounting rates based on information credibility. *RAIRO-Operations Research, 44*, 285–306.

Grant, A., Williams, P., Ward, N., & Basker, S. (2009). GPS jamming and the impact on maritime navigation. *The Journal of Navigation, 62*, 173–187.

Guo, H., Shi, W., & Deng, Y. (2006). Evaluating sensor reliability in classification problems based on evidence theory. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics), 36*, 970–981.

Hermans, F., Dziengel, N., & Schiller, J. (2009). Quality estimation based data fusion in wireless sensor networks. *IEEE 6th Int. Conf. on Mobile Adhoc and Sensor Systems* (pp. 1068–1070).

Iphar, C., Napoli, A., & Ray, C. (2015). Detection of false AIS messages for the improvement of maritime situational awareness. *OCEANS 2015-MTS/IEEE Washington* (pp. 1–7).

Jonker, C. M., & Treur, J. (1999). Formal analysis of models for the dynamics of trust based on experiences. *European Workshop on Modelling Autonomous Agents in a Multi-Agent World* (pp. 221–231).

Klir, G. J. (2005). *Uncertainty and information: foundations of generalized information theory*. John Wiley & Sons.

Laso, P. M., Brosset, D., & Puentes, J. (2017). Analysis of quality measurements to categorize anomalies in sensor systems. *Computing Conference* (pp. 1330–1338).

Lenart, M. (2018). Credibility evaluation in the context of event-type messages from sensors. *Schedae Informaticae, vol. 27, accepted*.

Lenart, M., Bielecki, A., Lesot, M.-J., Petrisor, T., & Revault d'Allonnes, A. (2018). Dynamic trust scoring of railway sensor information. *Proc. of the 17th Int. Conf. on Artificial Intelligence and Soft Computing ICAISC'18, Lecture Notes in Artificial Intelligence, vol. 10842* (pp. 579–591).

Lenart, M., Bielecki, A., Lesot, M.-J., Petrisor, T., & Revault d'Allonnes, A. (2019). Trust dynamics: a case-study on railway sensors. *Proc. of the 8th Int. Conf. on Sensor Networks, SENSORNETS'2019* (pp. 47–57). Scitepress.

Lesot, M.-J., Delavallade, T., Pichon, F., Akdag, H., Bouchon-Meunier, B., & Capet, P. (2011). Proposition of a semi-automatic possibilistic information scoring process. *Proc. of the 7th Conf. of the European Society for Fuzzy Logic and Technology EUSFLAT-2011 and LFA-2011* (pp. 949–956). Atlantis Press.

Lesot, M.-J., & Revault d'Allonnes, A. (2017). *Information quality and uncertainty*, 135–146. Springer International Publishing.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *J. Data and Information Quality, 1*, 2:1–2:22.

Martin, A., Molteno, T., & Parry, M. (2014). Measuring the performance of sensors that report uncertainty. *arXiv preprint arXiv:1411.4354*.

Mercier, D., Quost, B., & Denœux, T. (2008). Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion, 9*, 246–258.

Mui, L. (2002). *Computational models of trust and reputation: Agents, evolutionary games and social networks*. Doctoral dissertation, MIT.

Naumann, F. (2002). *Quality-driven query answering for integrated information systems*, vol. 2261. Springer Science & Business Media.

Petit, J., & Shladover, S. E. (2014). Potential cyberattacks on automated vehicles. *IEEE Trans. on Intelligent Transportation Systems, 16*, 546–556.

Pichon, F., Dubois, D., & Denoeux, T. (2012). Relevance and truthfulness in information correction and fusion. *Int. Jour. of Approximate Reasoning, 53*, 159 – 175.

Pichon, F., Labreuche, C., Duqueroie, B., & Delavallade, T. (2014). Multidimensional approach to reliability evaluation of information sources. In P. Capet and T. Delavallade (Eds.), *Information evaluation*, chapter 5, 129–160. Wiley.

Pon, R. K., & Cárdenas, A. F. (2005). Data quality inference. *Proc. of the 2nd Int. Workshop on Information quality in information systems* (pp. 105–111).

Revault d'Allonnes, A. (2014). An architecture for the evolution of trust: Definition and impact of the necessary dimensions of opinion making. In P. Capet and T. Delavallade (Eds.), *Information evaluation*, chapter 9, 261–294. Wiley.

Revault d'Allonnes, A., & Lesot, M.-J. (2014). Formalising information scoring in a multivalued logic framework. *Proc. of the 15th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, (IPMU 14). Part I* (pp. 314–324). Springer.

Rogova, G., Hadzagic, M., St-Hilaire, M. O., Florea, M. C., & Valin, P. (2013). Context-based information quality for sequential decision making. *IEEE Int. Multi-Disciplinary Conf. on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA* (pp. 16–21).

Rogova, G. L., & Bossé, É. (2010). Information quality in information fusion. *Proc. of the 13th Conf. on Information Fusion FUSION'10*.

Samet, A., Lefevre, E., & Yahia, S. B. (2014). Integration of extra-information for belief function theory conflict management problem through generic association rules. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 22*, 531–551.

Shafer, G. (1976). *A mathematical theory of evidence*, vol. 42. Princeton university press.

Shao, S., Guo, S., & Qiu, X. (2017). Distributed fault detection based on credibility and cooperation for WSNs in smart grids. *Sensors, 17*, 983.

Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. *Proc. of Int. Conf. on Information Retrieval Knowledge Management* (pp. 300–304).

Todoran, I.-G., Lecornu, L., Khenchaf, A., & Caillec, J.-M. L. (2015). A methodology to evaluate important dimensions of information quality in systems. *J. Data and Information Quality, 6*, 11:1–11:23.

Todoran, I.-G., Lecornu, L., Khenchaf, A., & Le Caillec, J.-M. (2013). Information quality evaluation in fusion systems. *Proc. of the 16th Int. Conf. on Information Fusion, FUSION'13* (pp. 906–913).

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*, 5–33.

Xiao, W., Tham, C. K., & Das, S. K. (2010). Collaborative sensing to improve information quality for target tracking in wireless sensor networks. *8th IEEE Int. Conf. on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 99–104).

Young, S., & Palmer, J. (2007). Pedigree and confidence: Issues in data credibility and reliability. *Proc. of the Int. Conf. on Information Fusion, FUSION'07*.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*, 338–353.

Zahedi, S., Ngai, E., Gelenbe, E., Mylaraswamy, D., & Srivastava, M. B. (2008). Information quality aware sensor network services. *42nd Asilomar Conference on Signals, Systems and Computers* (pp. 1155–1159).

Oświadczenie

Niniejszą pracę doktorską wykonałem osobiście i nie korzystałem z innych prac, nie napisanych w źródłach.

.....