RAILENIUM Test & Research Centre
IFSTTAR - Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux
SNCF Réseau
Université de Lille

THÈSE

présentée en vue d'obtenir le grade de

DOCTEUR

en

Automatique, Génie Informatique, Traitement du Signal et des Images

par

Ci LIANG

Doctorat délivré par l'université de Lille

Titre de la thèse :

Contributions to Risk Modeling and Analysis at Railway Level Crossings

Contributions à la modélisation et l'analyse de risque aux passages à niveau

Soutenue le 10 avril 2018 devant le jury d'examen :

Président	Pr. Adnane Boukamel	IRT RAILENIUM
Rapporteur	Pr. Walter Schön	Heudiasyc, Université de Technologie de Compiègne
Rapporteur	Dr-HDR. Louahdi Khoudour	Céréma de Toulouse
Examinateur	Dr. Anne Silla	VTT, Finland
Invité	Pr. Wei Zheng	Beijing Jiaotong University
Invitée	Mme Marie-Hélène Bonneau	UIC, Paris
Référent de thèse	M. Olivier Cazier	SNCF Réseau
Co-Directeur de thèse	Dr. El-Miloudi El-Koursi	IFSTTAR-COSYS/ESTAS
Directeur de thèse	Dr-HDR. Mohamed Ghazel	IFSTTAR-COSYS/ESTAS

Thèse préparée au Laboratoire d'Évaluation des Systèmes de Transports Automatisés et de leur Sécurité
IFSTTAR, COSYS/ESTAS, Lille-Villeneuve d'Ascq
École Doctorale Sciences pour l'ingénieur ED 072 - Université de Lille

Contributions to Risk Modeling and Analysis at Railway Level Crossings

Abstract:
Accidents at railway level crossings (LXs) often give rise to serious material and human damage and highly impact the reputation of railway safety. Although research on LX safety has been an area of great interest over the past decades, the causes of collisions that occur at LXs remain insufficiently understood. This PhD thesis deals with advanced quantitative risk analysis and modeling techniques with the aim to improve safety at LXs. The contributions of the work reported in this thesis are four-fold:
Firstly, a preliminary statistical analysis is performed. Namely, we analyze the impact of various factors (transport mode, geographical region and traffic moment) on the risk level at LXs quantitatively. Based on the obtained results, the main transport mode (motorized vehicle) causing LX accidents and the most risky regions are identified. Then, based on field experiments carried out at 12 LXs throughout France, thorough quantitative analysis of motorist behavior is performed to acquire the knowledge of motorist violation mechanisms causing train-car collisions. In this stage, the Phase Classification Analysis (PCA) concept is adopted to analyze motorist behavior with regard to three functional phases of the LX closure cycle, i.e., Ph2 "Red Flash and Sire", Ph3 "Barriers Coming Down" and Ph4 "Barriers Down". The motorist violation during the closure cycle is analyzed in terms of schedule factors, vehicle speed, Ph4 duration, LX location (namely railway station nearby or not) and road traffic density, respectively. Moreover, an advanced statistical accident prediction model was further developed. Such a model takes a variety of impacting factors into account, i.e., the average daily road traffic, the average daily railway traffic, the annual road accidents, the vertical road profile, the horizontal road alignment, the road width, the crossing length, the railway speed limit and the geographic region. The model validation phase we carried out has shown that the model allows for estimating accident frequency with a considerably high accuracy and has a more appropriate form compared with the existing models pertaining to LX accident prediction. Subsequently, an effective and comprehensive modeling framework based on Bayesian networks (BNs) for risk reasoning is proposed. It consists of a set of integral processes, namely risk scenario definition, real field data collection and processing, BN model establishment and model performance validation. The performance validation results indicate that our BN risk model has sound estimation performance and allows us to consider the outcomes of the model to be trustworthy. Based on the causal BN model, forward inference and reverse inference are achieved to make consequence prediction and cause diagnosis. Besides, influence strength analysis and sensitivity analysis were performed to determine the influence strength of causal factors on consequence factors and which causal factors the consequence factors are mostly sensitive to.
In summary, the aforementioned contributions are a direct response to the key knowledge gap regarding the mechanisms underlying LX accidents. The obtained results pave the way for effectively identifying the riskiest LXs, and determining appropriate practical design measures and targeted technical solutions, so as to improve LX safety.

Keywords : Level crossing safety, Railway safety, Train-car collisions, Risk analysis, Accident prediction, Statistical modeling, Bayesian network modeling.

Contributions à la modélisation et l'analyse de risque aux passages à niveau

Résumé :
Du point de vue de la sécurité, les passages à niveau (PN) sont considérés comme des points critiques pour le transport ferroviaire. Les accidents aux PN constituent une part majeure des accidents ferroviaires et impactent considérablement la réputation de la sécurité ferroviaire en général. Bien que la sécurité des PN ait été une thématique de recherche largement investie au cours des dernières décennies, les mécanismes relatifs aux collisions qui surviennent au niveau des PN demeurent insuffisamment maîtrisés. Ce travail de thèse porte sur la modélisation et l'analyse quantitative des risques dans le but d'améliorer la sécurité aux PN. Les contributions de ce travail peuvent être présentées selon quatre volets :
Sur le premier volet, une analyse statistique préliminaire est effectuée. En particulier, une analyse quantitative de l'impact de divers facteurs (mode de transport, région géographique et moment de trafic) sur le niveau de risque aux PN a été effectuée. Sur le deuxième volet et en se basant sur des données expérimentales fines obtenues à partir de l'instrumentation de 12 PN à différents endroits en France, une analyse quantitative du comportement des automobilistes est réalisée pour explorer les mécanismes de violation lors de la traversée d'un PN. Pour ce fait, nous avons adopté le concept d'analyse à base de phases (ACP) afin d'étudier le comportement des automobilistes en fonction des trois phases fonctionnelles du cycle de fermeture d'un PN. La violation des automobilistes pendant le cycle de fermeture est ainsi analysée en tenant compte de différents facteurs, dont : la vitesse d'approche du véhicule, la durée de fermeture du PN, de localisation de PN (à proximité ou non d'une gare) et la densité du trafic routier. Sur le troisième volet, nous proposons un modèle statistique multi-paramètres de prévision des accidents. Le modèle établi prend en compte le trafic ferroviaire quotidien moyen, l'accidentologie routière, le profil et l'alignement horizontal de la route, la largeur de la route, la longueur du passage à niveau, la vitesse ferroviaire limite et la région géographique. À travers la phase de validation, nous avons montré que le modèle développé est très fiable et offre une précision bien plus élevée par rapport aux modèles de référence de la littérature. Sur le dernier volet, nous proposons un cadre de modélisation global basé sur des réseaux Bayésiens (RB) pour le raisonnement causal. Ce cadre comporte un ensemble de processus intégrés permettant de définir les scénarios de risque tout en intégrant les données réelles. Les résultats de la phase de validation indiquent que ce modèle de risque présente de bonnes performances en termes d'estimation. Par ailleurs, sur la base du RB développé, des analyses causales directe et inverse sont ensuite effectuées pour quantifier les conséquences de différents facteurs de risque et, inversement, réaliser le diagnostic relatif à des scénarios à risque donnés. En outre, des analyses d'influence et de sensibilité ont été adoptées pour déterminer le niveau d'influence associé à des facteurs de risque relativement à des scénarios critiques donnés, et identifier les facteurs les plus impactants.
En résumé, les contributions de ce travail de thèse offrent une réponse directe à l'insuffisante maîtrise des divers mécanismes qui sous-tendent les accidents aux PN. En outre, les résultats obtenus ouvrent la voie vers l'identification des solutions techniques et en termes de conception en vue d'améliorer la sécurité des PN.

Mots clés : Sécurité aux passages à niveau, Sécurité ferroviaire, Collisions train-voiture, Analyse de risque, Prédiction d'accident, Modèles statistiques, Réseaux Bayésiens.

Publications

Journals:

1) C. Liang, M. Ghazel, O. Cazier, E. M. El Koursi. Developing accident prediction model for railway level crossings. *Safety Science* 101, 48-59, Elsevier, 2018.

2) C. Liang, M. Ghazel, O. Cazier, E. M. El Koursi. A new insight on the risky behavior of motorists at railway level crossings: An observational field study. *Accident Analysis and Prevention* 108, 181-188, Elsevier, 2018.

3) C. Liang, M. Ghazel. A risk assessment study on accidents at French level crossings using Bayesian belief networks. *International Journal of Injury Control and Safety Promotion*, Taylor & Francis, 2018. https://doi.org/10.1080/17457300.2017.1416480

4) C. Liang, M. Ghazel, O. Cazier, E. M. El Koursi. Analyzing risky behavior of motorists during the closure cycle of railway level crossings. *Safety Science*, Elsevier, 2018. https://doi.org/10.1016/j.ssci.2017.12.008

5) C. Liang, M. Ghazel, O. Cazier, E. M. El Koursi. Risk analysis on level crossings using a causal Bayesian network based approach. *Transportation Research Procedia* 25, 2172-2186, Elsevier, 2017.

Submitted:

6) C. Liang, M. Ghazel, O. Cazier. Statistical modeling and validation for accident prediction at railway level crossings. *Safety Science*, Elsevier, 2018.

7) C. Liang, M. Ghazel, O. Cazier, L. Bouillaut. Advanced model-based risk reasoning on automatic Level crossings. *IEEE Trans. on Intelligent Transportation Systems*, 2018.

Conferences:

8) C. Liang, M. Ghazel, E. M. El Koursi and O. Cazier. Statistical Analysis of Collisions at French Level Crossings. *Proceedings of the Third International Conference on Railway Technology: Research, Development and Maintenance - RAILWAYS 2016*, Avril 2016, Sardaigne, Italy.

9) C. Liang, M. Ghazel, O. Cazier, E. M. El Koursi. Risk analysis on level crossings using a causal Bayesian network based approach. *World Conference on Transport Research - WCTR 2016*, July 2016, Shanghai, China. (selected to the journal Transportation Research Procedia)

10) C. Liang, M. Ghazel, L. Bouillaut, O. Cazier, E. M. El Koursi. Bayesian Network Modeling Applied on Railway Level Crossing Safety. *International Conference Reliability, Safety and Security of Railway Systems: Modelling, Analysis, Verification and Certification - RSSRail 2017*, November 2017, Pistoia, Italy.

11) C. Liang, M. Ghazel, O. Cazier. Using Bayesian Networks for the Purpose of Risk Analysis at Railway Level Crossings. *15th IFAC Symposium on Control in Transportation Systems - CTS 2018*, June 2018, Savona, Italy.

Glossary of Acronyms and Terms

ADRT	Average Daily Railway Traffic
ADRV	Average Daily Road Vehicle
AADT	Annual Average Daily Traffic
AIC	Akaike's Information Criterion
AIS	Adaptive Importance Sampling
ALCAM	Australian Level Crossing Assessment Model
ANB	Augmented Naïve Bayes
ATP	Automatic Train Protection
AUC	Area Under the ROC Curve
BIC	Bayesian Information Criterion
BN	Bayesian Network
BNI-RR	Bayesian Network based Inference for Risk Reasoning
BS	Bayesian Search
CAK	CAusal Knowledge
CDF	Cumulative Distribution Function
CM	Corrected Moment
CPT	Conditional Probability Table
CSC	Causal Structural Constraints
DAG	Directed Acyclic Graph
DF	Degree of Freedom
DSS	Decision Support System
EC	Existence Constraint
EGS	Essential Graph Search
EM	Expectation Maximization
EPIS	Evidence Pre-propagation Importance Sampling
ETA	Event Tree Analysis
ETS	Event Trees
FC	Forbidden Constraint
FL	Fuzzy Logic
FPR	False Positive Rate
FTA	Fault Tree Analysis
FTS	Fault Trees

GOF	Goodness of Fit
GSPNs	Generalized Stochastic Petri Nets
GTT	Greedy Thick Thinning
IDs	Influence Diagrams
LL	Log-likelihood statistic
LX	Level Crossing
MCs	Markov Chains
MLE	Maximum Likelihood Estimation
MV	Motorized Vehicle
NB	Negative Binomial
NLS	Nonlinear Least-Square
NNs	Neural Networks
NP-hard	Non-deterministic Polynomial-hard problem
PB	Pedestrian or Bicycle
PCS	Pearson Chi-square Statistic
PDC	Potential Directed Constraint
PNs	Petri Nets
ROC	Receiver Operating Characteristic Curve
SAL0	Passages à Niveau à Signalisation Automatique Lumineuse à feux seuls (automated LXs with flashing lights but without barriers)
SAL2	Passages à Niveau à Signalisation Automatique Lumineuse à 2 Demi-Barrières (automated LXs with two half barriers and flashing lights)
SAL4	Passages à Niveau à Signalisation Automatique Lumineuse à 4 Demi-Barrières (automated LXs with four half barriers and flashing lights)
SPAD	Signal Passed at Danger
STD	Sensitivity Tornado Diagram
TAN	Tree Augmented Naïve Bayes
TPR	True Positive Rate
ZINB	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson
ZIP	U.S. Department of Transportation

.

Acknowledgments

The work reported in this thesis was achieved in the framework of "MORIPAN project: MOdèle de RIsque pour les PAssages à Niveau" within the Railenium Test and Research center, in cooperation with the National Society of French Railway Networks (SNCF Réseau) and the French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR). This PhD was prepared with COSYS/ESTAS team of IFSTTAR (Villeneuve-d'Ascq).

To be honest, this work could not been achieved without several persons' assistance, support and encouragement!

First and foremost, my special appreciation and thanks are undoubtedly delivered to my supervisor Dr. Mohamed Ghazel for his day-to-day constant support and guidance during the past three years. Without his help, this work would not be accomplished.

Likewise, I am deeply indebted and thoroughly grateful to my co-supervisor Mr. Olivier Cazier for his valuable suggestions and support about data and materials, and his active involvement in all technical phases of this work. Without his support, we could not obtain the first-hand field materials and the project could not be launched.

Moreover, I would like to definitely express my kind thanks to my co-supervisor Dr. El-Miloudi El-Koursi. He gave me valuable suggestions and support on my work, made careful review on the present dissertation and provided helpful comments.

Besides, I would like to thank my project manager Dr. Sébastien Lefebvre. He gave me a lot of support and help of my project and other aspects. With his help, my work could go smoothly.

Furthermore, I express my sincere appreciation to Pr. Adnane Boukamel for agreeing to be the chair of my thesis committee. I would like to thank all of you who agreed to be the referees of this thesis and allocated your valuable time in order to evaluate the quality of this work: Pr. Walter Schön and Dr-HDR. Louahdi Khoudour, for your examination of the thesis and their very helpful comments and suggestions. I would like to thank Pr. Wei zheng and Dr. Anne Silla for agreeing to be member of the jury and examining my work.

I would also like to express my thanks to Benoit GUYOT who is with SNCF Réseau and provided us a lot of support related to field data, and various members of IFSTTAR - ESTAS and LEOST Laboratories, particularly Joaquin (introducing this project to me). I have learned so much from all of you! I will fail in my duty if I do not acknowledge some

of my friends who helped me a lot during the past three years. I mention Rui, Sana, Baisi, Abderraouf, Ni, Liu, Zeting, Yuchen, Ji and many others.

Finally, I would like to thank all the people that supported me in many aspects during these three years, making this work possible!

Ci LIANG
Lille - January 8^{th} 2018.

I cannot find words to express my gratitude
to my mother, my father & my lover
to all my family
You raise me up to more than I can be!

The search for something permanent is one of the deepest of the instincts leading men to philosophy.

- -Bertrand Russell- -

Man cannot discover new oceans unless he has courage to lose sight of the shore.

- -André Gide- -

Courage is the ladder on which all the other virtues mount.

- -Clare Boothe Luce- -

Only those who are never look up at the starry sky, don't fall into a pit.

- -Georg Wilhelm Friedrich Hegel- -

In fate had, the most can see people of integrity.

- -William Shakespeare- -

Contents

List of Figures

List of Tables

Part I

PRELIMINARY INTRODUCTION

Chapter 1

Introduction

Sommaire

Overview

This dissertation presents a synthesis of research works, which were carried out as a fruit of my Ph.D. (2015-2018) accomplished with the RAILENIUM Test & Research Center and hosted by the COSYS/ESTAS team (Evaluation of Automated Transport Systems and their Safety) within IFSTTAR, the French Institute of Science and Technology for Transport, Development and Networks (Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux). This Ph.D. thesis has been prepared within the doctoral school SPI (Engineering Sciences) at Lille 1 university/Université Lille Nord de France. This thesis was supervised by Dr. Mohamed Ghazel, senior researcher with IFSTTAR-COSYS/ESTAS. The works presented in this dissertation devote to advanced modeling techniques and thorough risk analysis at railway level crossings. The intention of this research is two-fold: (1) responding to the key knowledge gap regarding hazard analysis at LXs, while identifying the potential risky factors causing LX accidents and their corresponding contribution degree. (2) paving the way for identifying practical measures and improvement recommendations to improving LX safety.

This chapter is structured as follows: Section 1.1 exposes general introduction of the research context. Then, the research motivation and objective is stated in section 1.2. Section 1.3 offers the main contributions of our research reported in the present dissertation. Finally, section 1.4 shows the organization and structure of the dissertation.

1.1 General Introduction

A level crossing (LX) is an intersection where a railway line intersects with a road or a path at the same level. In the literature, various terms to designate such intersection can be met: "level crossing", "grade crossing", "railway-highway crossing" or simple "crossing". In France, the fact that the LX is railway property upon which road users are given permission to cross, is the reason for legislation and rules giving right of way (priority) to trains [Read 2016]. Accidents at LXs often give rise to serious material and human damage and seriously hamper railway safety reputation, although the majority of accidents are caused by road user violations. As demonstrated by accident and incident statistics, level crossing (LX) safety is one of the most critical issues that railway stakeholders need to deal with [Ghazel 2009, Liu 2016, Mekki 2012]. In 2012, there were more than 118,000 LXs in the 28 countries of the European Union (E.U.), and 5 LXs per 10 line-km on average in the E.U. [ERA 2014]. Accidents at European LXs account for about one-third of the entire European railway accidents and result in more than 300 deaths every year [Liu 2016]. In some European countries, LX accidents account for up to 50% of railway accidents [Ghazel 2014]. As an example, in the UK, LXs account for 11.8 "fatalities and weighted injuries" per year, comprising 8.4% of the total system risk for the railway network [Silmon 2010]. In the U.S., highway grade crossing users represent about 30% of all rail-related fatalities. Although grade crossing collision rates have declined 80% in the past 20 years, over 15,000 highway users have been killed over that time period at grade crossings [Administration 2012]. In Australia, there were 49 collisions between trains and road vehicles at LXs in 2011, leading to 33 fatalities. It could be noted that, in this country, the problem is not limited to collisions between trains and vehicles, since from 2002 to 2012 there were 92 collisions between trains and pedestrians at LXs [Bureau 2012]. In Finland, there are about 4000 LXs along around 6000 km of track [Kallberg 2017]. In this country, unauthorized access of road users to railways is fairly easy, because only a small fraction of LXs are fenced [Silla 2012]. Moreover, most of LXs are not equipped with active warning devices and located on minor gravel roads [Seise 2010].

In France, the railway network shows more than 18,000 LXs for 30,000 km of railway lines, which are crossed daily by 16 million vehicles on average, and around 13,000 LXs show heavy road and railway traffic [SNCF Réseau 2011]. In 2013, 148 train/vehicle

collisions occurred at French LXs, giving rise to 29 deaths. In 2016, 111 train/vehicle collisions at French LXs led to 31 deaths [Plesse 2017]. This number was half the total number of collisions per year at LXs a decade ago, but still too large.

LX safety involves various aspects: technical elements, operational procedures, human factors and environmental considerations. Due to non-deterministic causes, complex operation background and the lack of thorough statistical analysis based on detailed accident/incident data, risk assessment of LXs remains a challenging task. In order to significantly reduce the number of accidents and their related consequences at LXs, it is crucial to carry out a series of thorough analyses and modeling to understand the potential reasons for these accidents and thus, enable the identification of practical design and improvement recommendations to prevent accidents at LXs.

1.2 Research Motivation and Objective

There are four main LX types in France [SNCF 2015], as shown in Fig. 1.1:

a) SAL4: Automated LX systems with four half barriers and flashing lights;

b) SAL2: Automated LX systems with two half barriers and flashing lights;

c) SAL0: Automated LX systems with flashing lights but without barriers;

d) Crossbuck LXs, without automatic signaling.

(a) SAL4

(b) SAL2

(c) SAL0

(d) Crossbuck LX

Fig. 1.1. Four types of LXs in France

As shown in Table 1.1, SAL2 (more than 10,000) is the most widely used type of LX in France. Moreover, more than 4,000 accidents at SAL2 LXs contributed most to the total number of accidents at LXs from 1974 to 2014.

Table 1.1. Accidents at different types of LXs in France from 1974 to 2014

Type of LX	Number	# Accident
SAL4	> 600	> 600
SAL2	> 10,000	> 4,000
SAL0	> 60	> 50
Crossbuck LX	> 3,000	> 700

LX accidents involve the following transport modes: 1) motorized vehicle (MV), 2) pedestrian and bicycle (PB). As illustrated in Fig. 1.2, motorized vehicle is the main transport mode causing LX accidents in France. Moreover, as the LX accident frequency caused by motorized vehicles increases, the entire LX accident frequency increases accordingly. On the contrary, pedestrians and cyclists contribute very little to the overall risk of LX accidents [Liang 2017c].

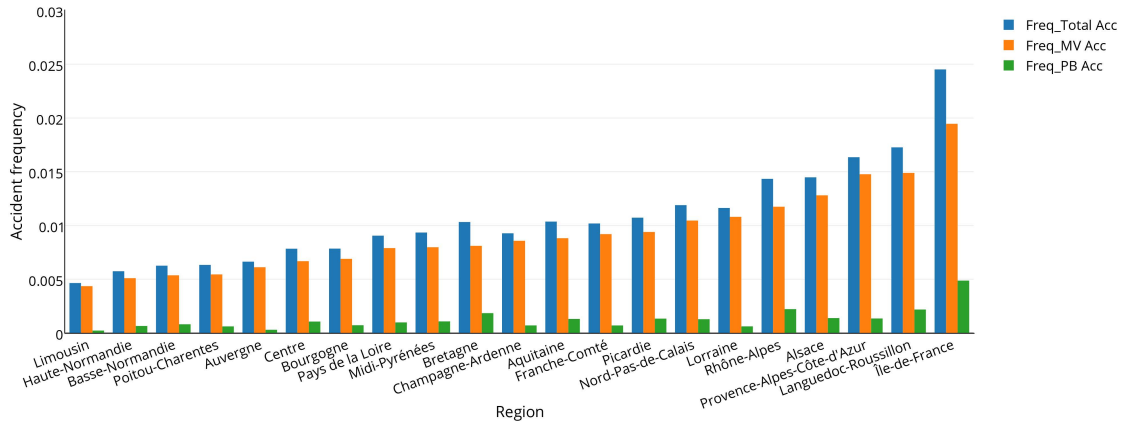


Fig. 1.2. Comprehensive accident frequency in different regions

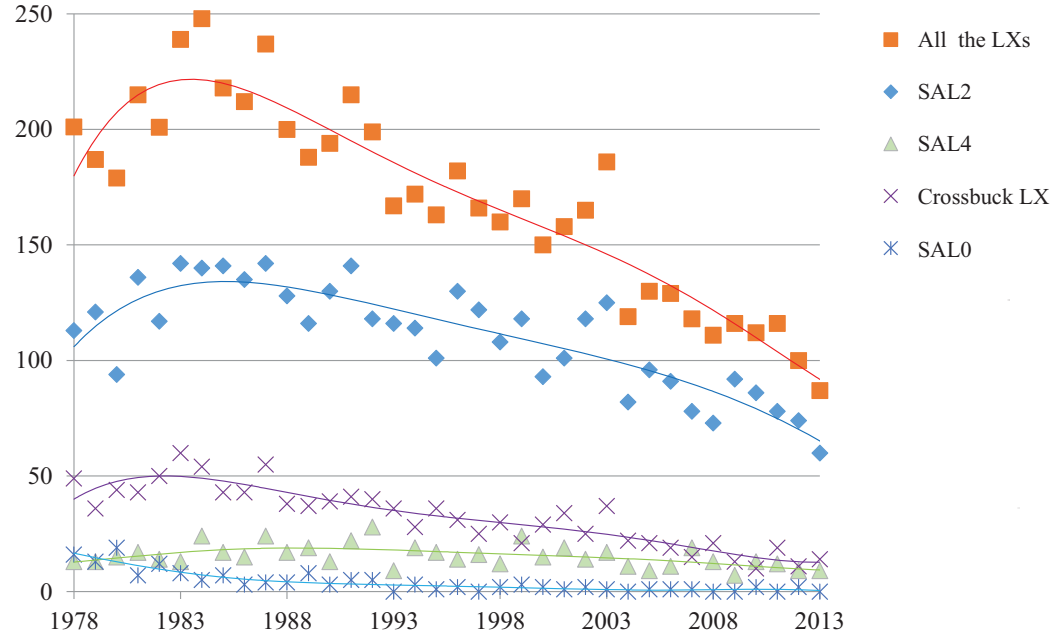


Fig. 1.3. The number of collisions (train-MV) at different types of LX in France from 1978 to 2013

Considering the train/motorized vehicle (train-MV) collisions, SAL2 LXs also have the major part of LX accidents according to the statistics shown in Fig. 1.3. In fact, the goal of SNCF is to perform analysis that paves the way toward improving significantly LX safety as a whole. Hence, being given the number of SAL2 LXs, dealing with this LX category consists a priority issue for SNCF. In addition, according to the SNCF statistics, the accidents at SAL2 LXs can be considered as the most representative for LX accidents

in general. For all these reasons, in agreement with SNCF, in the scope of this work, we focus on the analysis of train-MV collisions occurring at SAL2 LXs while exploiting the results, as much as possible, to desire general conclusions.

It should be noted that suicide scenarios are not in the scope of our global study. In addition, in this study, the LX control system failure is not considered, since this is a purely technological problem that can be addressed in a different way using classical techniques. Besides, it is worth recalling that less than 1% of accidents are linked to a dysfunction of LX control system.

1.3 Main Contributions

This dissertation focuses on advanced risk analysis on LXs from the four aspects of preliminary analysis on potential influential factors, accident frequency prediction modeling, motorist behavior quantitative analysis and Bayesian Network (BN) risk model. The main contributions are all discussed in the second part (Part II) of the thesis and can be presented as follows:

Preliminary statistical analysis on LXs

In this stage, various kinds of impacting factors, namely transport mode, geographical region and traffic moment, are analyzed by means of statistical techniques to dig out their statistical characteristics based on the accident data from SNCF, the French national railway operator. Then, we assess the effect of various factors on the risk level quantitatively, in such a way as to open the way for setting efficient solutions and consequently, reaching the point of improving LX safety. In details:

1. A general risk analysis of average accident frequency in terms of transport mode and geographical region is performed.

2. Then, the risk analysis in terms of frequency coefficient, namely the average accident frequency acted by the traffic moment, is performed with regard to various traffic moment groups.

3. The frequency coefficient distributed in different French regions is generated.

Quantitative analysis of motorist behavior

Collisions between trains and motorized vehicles contribute most to LX accidents, while the risky behavior of motorists is the primary cause of such accidents. Therefore, in order to acquire a better understanding of risky motorist behavior at LXs, a risk analysis of motorist behavior is performed based on field measurement conducted at 12 automatic LXs (11 equipped with two half barriers (SAL2) and 1 equipped with four half barriers

(SAL4)). We focus on motorist behavior during the LX closure cycle while distinguishing between different phases. Namely, the closure cycle is divided into three phases which are "Ph2 Red Flash and Siren", "Ph3 Barriers Coming Down" and "Ph4 Barriers Down". This gives a novel insight compared with existing studies which analyze motorist behavior during a whole mixed LX control cycle. A thorough statistical analysis is subsequently performed according to the phase periods. In what follows, we give the detailed description of the related contributions.

1. Global violation trend of motorist behavior during Ph2 and Ph3 is investigated firstly.

2. An analysis on the violation rate during Ph2 and Ph3 according to the week and the hour is performed.

3. An analysis on the speed of violating vehicles during Ph2 and Ph3 is performed.

4. We analyze the impact of prolonged Ph4 duration on the zigzag violation rate of motorists.

5. The impact of LX location (near to railway station or not) on the zigzag violation rate of motorists during Ph4 is analyzed.

6. The impact of road traffic density on the waiting queue in front of LXs and troop crossing phenomenon is investigated.

7. The comparison of motorist responses to SAL2 and SAL4 LXs is performed.

Advanced statistical accident prediction modeling

In this stage, we develop a new statistical model to predict accident frequency at LXs. The detailed contributions are shown as follows:

1. An advanced accident frequency prediction model, which enables to rank LXs accurately based on the impacting parameters and identify the significant impacting parameters efficiently, is developed.

2. In this model, we take the influential parameters into account, namely, the average daily road traffic, the average daily railway traffic, the annual road accidents, the vertical road profile, the horizontal road alignment, the road width, the crossing length, the railway speed limit and the geographic region. We will better explain these factors in the sequel.

3. The Nonlinear Least-Squares (NLS) method, Poisson regression method, negative binomial (NB) regression method, zero-inflated Poisson (ZIP) regression method and zero-inflated negative binomial (ZINB) regression method are employed to estimate the respective coefficients of parameters in the prediction model.

4. Then, a validation process is performed based on various statistical and probabilistic means to examine how well the estimation of the model fits the reality.

5. Finally, a comparison between the present model and two existing reference models is carried out to assess the efficiency of our model.

Bayesian network based framework for LX risk reasoning

In this stage, we adopt the Bayesian network (BN) to develop a risk model for LX accident/consequence estimation and causality diagnosis. Based on the investigation of some existing modeling techniques (refer to chapter 2), BN technique is considered as the appropriate notation due to its high computational efficiency, outstanding advantages involving the conjunction of domain expertise and automatic structure/parameters learning and, most importantly, causality analysis based on both forward inference (deductive reasoning) and reverse inference (abductive reasoning) [Weber 2012]. In what follows, we give the detailed contributions.

1. A BN based framework for causal reasoning related to risk analysis is proposed. It consists of a set of integrated stages, namely, risk scenario definition, real field data collection and processing, BN model establishment and model performance validation.

2. Causal structural constraints are introduced to the framework for the purpose of combining empirical knowledge with automatic learning approaches, thus to identify effective causalities and avoid inappropriate structural connections.

3. The proposed framework is applied to risk analysis of LX accidents in France. Namely, the BN risk model is established based on the real field data of LX accidents/incidents and the model performance is evaluated.

4. Forward inference and reverse inference based on the BN risk model are performed to predict LX accident occurrence and quantify the contribution degree of various impacting factors respectively, so as to identify the riskiest factors.

5. Influence strength and sensitivity analyses are further carried out to scrutinize the influence strength of various causal factors on the LX accident occurrence and determine which factors the LX accident occurrence is most sensitive to.

1.4 Organization and Structure of the Dissertation

This dissertation is organized in three parts and the chapters in each part are structured as follows:

PART I: This part is the precursor of the present dissertation, which is composed of two chapters:

Chapter 1 - Introduction: we give the introduction of the research context and present the motivations and the main contributions of our work.

Chapter 2 - Literature Review: a literature review regarding LX safety analysis, statistical accident prediction and risk analysis modeling is performed.

PART II: This part is dedicated to the main contributions of our research, and is composed of four chapters:

Chapter 3 - Preliminary statistical analysis on LXs

Chapter 4 - Motorist Behavior Quantitative Analysis: Experiments at 12 selected automated LXs

Chapter 5 - Advanced statistical accident prediction modeling

Chapter 6 - Bayesian network based framework for LX risk reasoning

Please refer to section 1.3 for the detailed works in each chapter of this part.

PART III: This part discusses conclusion remarks regarding the dissertation and the future research directions, and is composed of one chapter:

Chapter 7 - Overall Conclusions And Perspectives

Chapter 2

Literature Review

Sommaire

Rail transportation is one of most important modes of transport throughout the world. Hundreds of millions of persons use this transportation mode every day, and an important share of goods is carried by train.

Railways involve a complex operation background, safety–critical control systems and various stakeholders. Moreover, due to the increasingly strong needs of high speed and high carrying capacity, railway safety continues to be one of the most critical issues that railway stakeholders need to deal with.

Railway accidents refer to accidents that affect normal railway operation safety, such as train-train collisions, derailments, fires, explosions and power failures during the operation of trains as well as collisions with pedestrians, motorized vehicles, non-motorized vehicles, livestock and other obstructions during the operation of the train, and even serious late delays caused by improper management and operation. In the E.U., railway passengers have lower traveling risks (0.156 fatalities per billion passenger-kilometers) in comparison to other means of land transportation mode such as buses (0.433 fatalities per billion passenger-kilometers), cars (4.450 fatalities per billion passenger-kilometers) and motorcycles (52.593 fatalities per billion passenger-kilometers) [E.U. 2012]. In 2014, the E.U. member states reported 2,076 significant railway accidents resulting in 1,054 fatalities and 819 serious injuries. This represents a 5 % increase in the number of significant accidents and a 7 % drop in casualties compared to 2013 [ERA 2016]. The level of railway safety is traditionally expressed as the accident and casualty risk being a rate of the number of outcomes per exposure. Considering all railway fatalities (excluding suicides), the fatality risk per million train-km in the period 2010-2014 was 0.28 killed per million train-kilometers at the E.U. level. The fatality risk of railway passengers was 0.14 killed

passengers per billion train-kilometers in the period 2006-2014. Although the global safety performance of railways in E.U. member states is high, it appears that the safety levels vary greatly among member states and serious accidents continue to occur.

In the past, train-train collisions (internal accidents) due to signal passed at danger (SPAD) were the main accident scenario in railways, but since automatic train protection (ATP) systems were installed, the number of such accidents has been dramatically decreased. Nowadays, accidents at LXs (external accidents) represent one of the major concerns for railways and draw much attention of stakeholders and researchers. For instance, very recently on December 14, 2017, a train and a school bus collided at an automatic SAL2 LX near Perpignan in southern France, killing six children between eight and 14 years old and injuring more than 20 others [Willsher 2017]. This accident happened at a SAL2 LX on a two-lane road as the bus crossed a single-track railway line secured by a barrier and warning lights in each direction. In the following content, we will focus on the review of related works on LX safety and approaches pertaining to improving LX safety.

2.1 Related Works on Railway Level Crossing Safety

In the literature, a number of works have dealt with LX safety. In early studies, vehicle driver behavior with respect to the warning time of LX control was studied by [Richards 1990]. This study indicated that an increase in warning time was directly linked with an increase in risky behavior. [Ward 1995] reported a field study regarding a signage warning system for passive LXs with restricted lateral visibility. The proposed signage incorporates a stop sign and an advance warning to announce the restricted visibility. In this study, the vehicles were tracked through a series of seven sonar units placed along the roadside at incremental distances from the LX. It was reported that the approach speed of the vehicles approaching the LX did decrease after the installation of the signs. [Abraham 1998] have studied driver behavior at 37 LXs in Michigan and reported various driver characteristics and vehicle characteristics to be important safety factors. This study showed that the motorists approaching a multi-track LX from a multi-lane approach commit more violations. [Carlson 1999] developed two logistic regression models to predict whether drivers will commit violations at gated LXs. Such models can be used to identify gated LXs that are expected to have high violation rates compared with other gated LXs. [Cooper 2007] presented a research on selecting countermeasure improvements for LXs in California and suggested the use of two-long-arm gates or standard two-half gates with an additional median separator for LX safety improvement according to the calculated cost/benefit ratio. [Jia 2007] analyzed the impacts of different daily periods, seasons and railway lines on the probability of accidents occurring at LXs; but some kinds

of important factors, such as traffic moment, transport mode and geographical region were not taken into account. [McCollister 2007] investigated daily periods, train speeds and environmental factors at American LXs and proposed a statistical model to predict the probability of accidents. However, the predicted results have certain deviation with the actual situation due to the lack of thorough statistical analysis of the potential relationship between important impacting factors and accidents. [Hao 2013] identified the contributing factors that influence the severity of injuries in accidents at U.S. LXs. This study showed that the peak hour, vehicle speed, annual average daily traffic, train speed, area type were significant. [Khattak 2009] has made a comparison of vehicle driver behavior at LXs in two cities of Nebraska. This research has showed that drivers' unsafe behavior at LXs is location-dependent, but the magnitude of reduction in unsafe driver actions as a result of the installation of a separator along the road center at different LXs is more or less similar.

Although research on human factors related to LX safety has been an area of great interest over the past decades [Wilson 2005, Wilson 2014], the causes of collisions that occur at LXs remain insufficiently understood. In general, the causes of such collisions fall into two broad classes, which are unintentional error and intentional violation. For instance, a study presented by [Salmon 2013] describes a collision between a loaded semi-trailer truck and a train, which occurred in North Victoria, Australia. In this study, according to the investigation of the Office of the Chief Investigator (OCI) and court, the flashing lights and warning bells failed to alert the truck driver about the presence of the train early enough to initiate a braking action. As a consequence, the truck driver was not aware of the train approaching until it was too late to stop (unintentional error). The study conducted by [Davey 2008] analyzes the intentional violation of heavy vehicle drivers when they cross LXs. This study particularly discusses vehicle driver complacency due to the high level of familiarity. As for the unintentional errors, vehicle drivers may, for instance, fail to observe the warnings or fail to determine the braking distance appropriately. However, as for the intentional violations, vehicle drivers observe the warnings and fully understand their meaning, but intentionally commit transgressions on their own judgment [Lenné 2011].

In recent years, some available studies adopted qualitative approaches for understanding vehicle driver behavior during the entire cycle of LX control [Shappel 2000, San Kim 2013]. For example, a systems analysis framework [Leveson 2011, Read 2016, Wilson 2014] and a psychological schema theory [Stanton 2011, Salmon 2013] have been used to analyze the contributory factors behind LX accidents. [Salmon 2016] adopted the Cognitive Work Analysis (CWA), which is a systems approach with theoretical underpinnings in system theory, to achieve comprehensive understanding of different functions, decisions, strategies and tasks relevant to various LX stakeholders (road user, rail user and author-

ities). This study suggested further mechanisms for collecting and analyzing actual field data, which are crucial for allowing stakeholders to identify the key design requirements for improving safety at LXs. Obviously, these qualitative analyses are limited to quantify the frequency of accident occurrence and the contribution degree of impacting factors to accident occurrence.

On the other hand, some recent works (involving qualitative and quantitative means) employ surveys [Wigglesworth 2001], interviews [Read 2016], focus group methods [Stefanova 2015] or driving simulators [Larue 2015], rather than collecting real field data. For example, [Davey 2008] performed a study that aims to explore the contributing factors toward collisions at LXs from the perception of both heavy vehicle drivers and train drivers. In particular, a survey was conducted and samples of train drivers and heavy vehicle drivers were recruited as a series of focus groups. In this study, the contributing factors to heavy vehicle accidents at LXs such as the size of heavy vehicles, the carriage length, the line of sight and the impeded acceleration were assessed. [Lenné 2011] examined the effect of installing active controls, flashing lights and traffic signals on vehicle driver behavior. This quantitative study was achieved through the driving simulation. [Rudin-Brown 2012] also carried out a driving simulator based study that compares the efficacy and drivers' subjective perception of two active LX traffic control devices: flashing lights with boom barriers and standard traffic lights. A study launched by [Young 2015] examined where drivers direct their attention on approach to urban LXs located in busy shopping strip areas in Melbourne, and whether this differs between novice and experienced drivers. Instrumented vehicles, driver Verbal Protocols and post-drive Critical Decision interviews were utilized in this study. [Tey 2011] conducted an experiment to measure vehicle driver response to LXs equipped with stop signs (passive), flashing lights and half barriers with flashing lights (active), respectively. In this study, both a field survey and a driving simulator have been utilized. However, we believe that the reaction of vehicle drivers in simulation context potentially differs from that in reality, due to the different levels of feeling of danger. Although among these recent works, there were some quantitative analyses of vehicle driver behavior at LXs, few of them were carried out with regard to the separate phases of the automated LX closure cycle. Namely, these studies investigated vehicle driver behavior during the whole control cycle of LX, including when the LX is open to road traffic (theoretically, there is no risk for crossing LXs during the opening cycle).

2.2 Related Works on Accident Prediction Statistical Model

A number of existing works dealing with LX safety are devoted to developing qualitative approaches, in order to understand the potential reasons causing accidents at LXs [Davey 2008, Larue 2015, Read 2016, Stanton 2011, Stefanova 2015, Tey 2011, Wigglesworth 2001]. Although these available qualitative approaches are beneficial to understand the factors causing LX accidents, they do not allow for quantitatively predicting the likelihood of accident occurrence, or quantifying the contribution degree of the various impacting factors. Thereby, quantitative analysis approaches are crucial to thoroughly investigate the impacting factors and enable the identification of practical design and improvement recommendations to prevent accidents at LXs.

2.2.1 Peabody-Dimmick Formula

Some previous quantitative studies on prediction models to estimate the frequency of accidents or violations at LXs open a significant view on understanding the LX risk mechanism. In 1941, L. E. Peabody and T. B. Dimmick developed Peabody-Dimmick Formula [Ogden 2007] which is one of the earliest railway-highway crossing accident prediction models to estimate the number of accidents at railway-highway crossings in 5 years. As shown in Eq. (2.1), the parameters considered in this formula are the average daily road traffic V, the average daily railway traffic T and the protection coefficient indicative of warning devices adopted P, while K is an additional parameter. This formula was developed based on the accident data of rural railway-highway crossings in 29 states in the U.S. and was utilized through the 1950s. However, advances in both warning device technologies and LX design features quickly led to a no longer applicability of the predefined formula form and coefficients that reflect the conditions pertaining to LX accidents in 1941.

$$A_5 = \frac{1.28 \times (V^{0.170} \times T^{0.151})}{P^{0.171}} + K \tag{2.1}$$

2.2.2 New Hampshire Index

Subsequently, an evolutionary model of LX accident prediction called the New Hampshire Index [Oh 2006] was developed, which is given as follows:

$$HI = V \times T \times P_f \tag{2.2}$$

where HI represents the hazard index; V is the average daily road traffic; T is the average daily railway traffic and P_f is the protection factor indicative of the warning devices

adopted ($P_f = 0.1$ for automatic gates, $P_f = 0.6$ for flashing lights, $P_f = 1.0$ for signs only).

The New Hampshire model is a relative formula which can be used to rank the importance of crossing upgrades.

Several modifications of the New Hampshire Index are in use. In particular, some states in the U.S. use various other values for P_f as follows:

- $P_f = 0.13$ or 0.10 for automatic gates.
- $P_f = 0.33$, 0.20 or 0.60 for flashing lights.
- $P_f = 0.67$ for wigwags .
- $P_f = 0.50$ for traffic signal preemption.
- $P_f = 1.00$ for crossbucks.

Due to the simplicity of the New Hampshire Index, it has been widely used across the U.S. However, it is limited in that it does not predict the expected number of collisions, but only gives some indications about the priorities in terms of LX safety.

2.2.3 USDOT Formula

An accident prediction formula developed by the U.S. Department of Transportation (USDOT) in the early 1980s sought to overcome the limitations of earlier models [Chadwick 2014]. This comprehensive formula comprises three primary equations:

$$a = K \times EI \times MT \times DT \times HP \times MS \times HT \times HL \tag{2.3}$$

$$B = \frac{T_0}{T_0 + T} \times a + \frac{T}{T_0 + T} \times (\frac{N}{T}),\ T_0 = \frac{1}{0.05 + a} \tag{2.4}$$

$$A = \begin{cases} 0.7159 \times B, \text{for passive devices;} \\ 0.5292 \times B, \text{for flashing lights;} \\ 0.4921 \times B, \text{for gates;} \end{cases} \tag{2.5}$$

where a is the initial collision prediction (prediction of collisions per year at a given LX); K is the formula constant; EI is the exposure index (a variant of traffic moment) based on the product of highway and railway traffic; MT is the index for the number of main tracks; DT is the index for daily through trains during daylight; HP is the index for highway paved; MS is the index for maximum train speed; HT is the index for highway type and HL is the index for highway lanes. B is the adjusted accident frequency; T_0 is

the weighting factor and N is the number of accidents observed in T years at a given LX. Finally, A is the normalized accident frequency.

The USDOT formula is the most commonly used model in the U.S. today. As shown in Table 2.1, a specified table of USDOT provides each of the indexes for LXs equipped with passive controls, flashing lights and gates [Austin 2002]. Although the formula is comprehensive, its current definition makes it difficult to identify or prioritize the design or improvement activities that would most effectively address LX safety-related issues, since it does not provide the magnitude of characteristics' contribution to the LX safety.

Table 2.1. Indexes for USDOT Accident Prediction Formula

Index	Coefficient/relationship for passive control	Coefficient/relationship for flashing lights	Coefficient/relationship for automatic gates
K	0.002268	0.003646	0.001088
EI	$((ct+0.2)/0.2)^{0.3334}$	$((ct+0.2)/0.2)^{0.2953}$	$((ct+0.2)/0.2)^{0.3116}$
DT	$((d+0.2)/0.2)^{0.1336}$	$((d+0.2)/0.2)^{0.0470}$	1.0
MS	$e^{0.0077ms}$	1.0	1.0
MT	$e^{0.2094mt}$	$e^{0.1088mt}$	$e^{0.2912mt}$
HP	$e^{-0.6160(hp-1)}$	1.0	1.0
HL	1.0	$e^{0.1380(hl-1)}$	$e^{0.1036(hl-1)}$
HT	$e^{-0.1000(ht-1)}$	1.0	1.0

c=number of highway vehicles per day, t=number of trains per day, mt=number of main tracks, d=number of through trains per day during daylight, hp=highway paved (yes=1.0 and no=2.0), ms=maximum timetable speed in mph,hl=number of highway lanes, ht=highway type factor (defined as urban and rural, 1=interstate,..., 6=local) [Schoppert 1968].

2.2.4 ALCAM Model

In Australia, a model called Australian Level Crossing Assessment Model (ALCAM) was developed. It is a location specific and parameterized risk model which provides a method for assessing the risk level to LX users, train passengers and train staff [Woods 2008]. The ALCAM model is given as follows:

$$\text{ALCAM Risk Score} = \text{Infrastructure Factor} \times \text{Exposure Factor} \times \text{Consequence Factor} \tag{2.6}$$

where the Infrastructure Factor is the output of a complex scoring algorithm that assesses how the physical properties at each LX site would affect human behavior; the Exposure Factor is a function of the LX control type, vehicle (or pedestrian) volumes and train

volumes (i.e., the Peabody-Dimmick Formula is used as the Exposure Factor function) to address the combined exposure of trains and road vehicles (or pedestrians) pertaining to various LX control types; the Consequence Factor is the expected consequence of a collision which includes deaths and injuries involving both railway and roadway. The Infrastructure Factor adjusts the accident probability per year to reflect the actual LX site conditions. Multiplying the Infrastructure Factor by the Exposure Factor gives the actual annual likelihood of an accident occurring at a given LX [Committee 2012]. The Consequence Factor is expressed in terms of an expected number of equivalent fatalities per year. An equivalent fatality is a combination of all types of harm using the ratio: 1 fatality = 10 serious injuries = 200 minor injuries.

The output of ALCAM model enables to allocate weights to characteristics and controls at a level crossing to calculate a Likelihood Factor. The weightings applied have been determined through a series of workshops with contribution from experts including representatives from each mainland state of Australia and New Zealand covering expertise in road and rail engineering.

The ALCAM has been applied across all Australian states and in New Zealand since 2003, and overseen by a committee of representatives from the various jurisdictions of these countries to ensure its consistency in terms of development and application. However, the ALCAM does not cover all kinds of LX accidents, since its main focus is deliberate and accidental collisions involving user errors but excluding vandalism and suicide. It should be noticed that some LX physical properties considered in ALCAM show a high correlation between each other, which implies the existence of a kind of redundancy between the model inputs, and consequently a bias in terms of the outputs.

2.2.5 Poisson regression Model and its variants

In recent years, Poisson regression model and its variants, for instance, negative binomial (NB) regression model, zero-inflated Poisson (ZIP) regression model and zero-inflated negative binomial (ZINB) regression model [Lord 2010, Guikema 2012] have been mostly preferred to deal with risk/accident statistical analysis.

The Poisson regression model shown as Eq. (2.7) is the natural choice for modeling accident occurrence.

$$Poi(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, \ldots \tag{2.7}$$

where $Poi(X = k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$ and λ is the expectation value of the number of accidents.

The relationship between the mean value and the variance in the Poisson model is

given as follows:

$$VAR(X) = E(X) \tag{2.8}$$

However, [Chang 2005] indicates that accident frequency is likely to be over-dispersed (cf. Eq. (2.9)) and suggests using the negative binomial (NB) regression model as an alternative to the Poisson model.

$$VAR(X) \begin{cases} = E(X) \\ > E(X), \text{ over-dispersed} \\ < E(X), \text{ under-dispersed} \end{cases} \tag{2.9}$$

The NB model as a special case of Poisson-Gamma mixture model is a variant of the Poisson model designed to deal with over-dispersed data [Buddhavarapu 2016, Lord 2010, Utkin 2015]. The over-dispersion could come from several possible sources, e.g., omitted variables, uncertainty in exposure data, covariates or non-homogeneous LX environment [Miaou 1994]. The NB model considered in this study has the following expression:

$$P_{NB}(X = k) = \frac{\Gamma\left(k + \frac{1}{\alpha}\right)}{\Gamma\left(k+1\right)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k}, \; k = 0, 1, 2, \ldots \tag{2.10}$$

where $P_{NB}(X = k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$; λ is the expectation value of the number of accidents and α is the dispersion parameter.
The relationship between the mean value and the variance in the NB model is given as follows:

$$VAR(X) = E(X) + \alpha E(X)^2 \tag{2.11}$$

If $\alpha > 0$, there is an over-dispersion; if $\alpha < 0$, there is an under-dispersion and in the case where $\alpha = 0$, the NB model reduces to the Poisson model.

In practice, the count data may contain extra zeros relative to the Poisson or NB distribution. In this case, the ZIP or ZINB regression model is useful for analyzing such data [Ridout 2001]. The ZIP model is expressed as follows:

$$P_{ZIP}(X = k) = \begin{cases} \omega + (1-\omega)exp(-\lambda), \; k = 0; \\ (1-\omega)exp(-\lambda)\lambda^k/k!, \; k > 0 \end{cases} \tag{2.12}$$

where $P_{ZIP}(X = k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$; λ is the expectation value of the number of accidents and $log\left(\frac{\omega}{1-\omega}\right) = z'\gamma$ is the ZI link function that z' is the ZI covariate and γ is the corresponding ZI coefficient.

The mean value and variance of ZIP model are expressed as follows:

$$\begin{aligned} E(X) &= (1-\omega)\lambda \\ VAR(X) &= (1-\omega)\lambda(1+\omega\lambda) \end{aligned} \tag{2.13}$$

The ZINB model is expressed as follows:

$$P_{ZINB}(X=k) = \begin{cases} \omega + (1-\omega)(1+\alpha\lambda)^{-1/\alpha}, & k=0; \\ (1-\omega)\dfrac{\Gamma\left(k+\frac{1}{\alpha}\right)}{\Gamma(k+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha}\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k}, & k>0 \end{cases} \tag{2.14}$$

where $P_{ZINB}(X=k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$; λ is the expectation value of the number of accidents.

The mean value and variance of ZINB model are expressed as follows:

$$\begin{aligned} E(X) &= (1-\omega)\lambda \\ VAR(X) &= (1-\omega)\lambda(1+\omega\lambda+\alpha\lambda) \end{aligned} \tag{2.15}$$

The ZINB reduces to the ZIP in the limit $\alpha \to 0$.

[Oh 2006] adopted the expressions of the estimated expectation value $\hat{\lambda}$ as shown in Eq. (2.16) corresponding to the Poisson regression and NB regression models respectively when developing the U.S. LX accident prediction model. [Medina 2015] compared the USDOT model with the ZINB model, in terms of accident prediction accuracy, using the LX accident data from Illinois. The results of this study shows that the ZINB model has higher accuracy of prediction. [Lu 2016] employed the ZIP model to deal with LX accident prediction involving the data in North Dakota. It should be noted that the expressions of estimated $\hat{\lambda}$ as shown in Eq. (2.16) are not appropriate in our current study, since they are limited to handling zero observations and some impacting variables should not be in the exponential form (e.g., there is a logical assumption that the predicted LX accident frequency should be 0 if the average daily railway traffic is 0).

$$\begin{aligned} \hat{\lambda}_{Poi} &= exp\left(\sum_{j=1}^{m}\beta_0 + \beta_j x_j\right) \\ \hat{\lambda}_{NB} &= exp\left(\sum_{j=1}^{m}\beta_0 + \beta_j x_j + \varepsilon\right) \end{aligned} \tag{2.16}$$

where β is the estimated regression coefficient; x is the impacting variable and ε is the gamma-distributed error in the NB regression model.

[Miranda-Moreno 2005] developed another model of $\hat{\lambda}$ as shown in Eq. (2.17). In this model, the product of the average daily road traffic V and the average daily railway traffic T (known as the conventional traffic moment [Liang 2017c]) is adopted. As reported in [Khattak 2012], the combined exposure of V and T is more suitable than single V

for LX collision prediction, because the probability of train-involved LX collisions in the absence of trains is zero. However, since the road traffic density and railway traffic density have different levels of impact on accident occurrence in practice, the same power of V and T in Eq. (2.17) may cause relatively high deviation from reality.

$$\hat{\lambda} = (V \times T)^{\beta_1} exp\left(\sum_{j=1}^{m} \beta_j x_j + \sigma\right) \tag{2.17}$$

where $\sigma = \beta_0$ in Poisson regression model or $\sigma = \beta_0 + \varepsilon$ in NB regression model.

Based on these aforementioned investigations, it is clear that there is a strong need for an appropriate accident prediction model that should be comprehensive in terms of contributing factors to LX safety. Moreover, such a model should have high predictive accuracy.

2.3 Related Works on causal Modeling

Nowadays, risk analysis approaches are required to deal with increasingly complex systems with a large number of involved parameters. Moreover, an intelligent decision support system for risk analysis shall have the ability of making inference based on the risk causal knowledge. Therefore, such approaches should fulfill the following characteristics:

- Having strong modeling ability,
- Having high computational efficiency,
- Providing simple means to specify risk scenarios,
- Offering effective reasoning between risky factors and scenarios,
- Effectively identifying the most important risky factors.

In the domain of risk analysis, various approaches are adopted for the modeling and analysis process. In this section, we introduce some main relevant formalisms and approaches, namely the Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Markov Chains (MCs), Petri Nets (PNs), Fuzzy Logic (FL), and Bayesian Networks (BNs).

2.3.1 Fault Tree Analysis

Due to the combination of qualitative and quantitative analysis, the Fault Tree Analysis (FTA) developed by H.A. Watson at Bell Laboratories [Ericson 1999] has been widely used for risk analysis in various contexts. FTA is a deductive and top-down method which aims at analyzing the effects of initiating faults and events on a complex system and offers

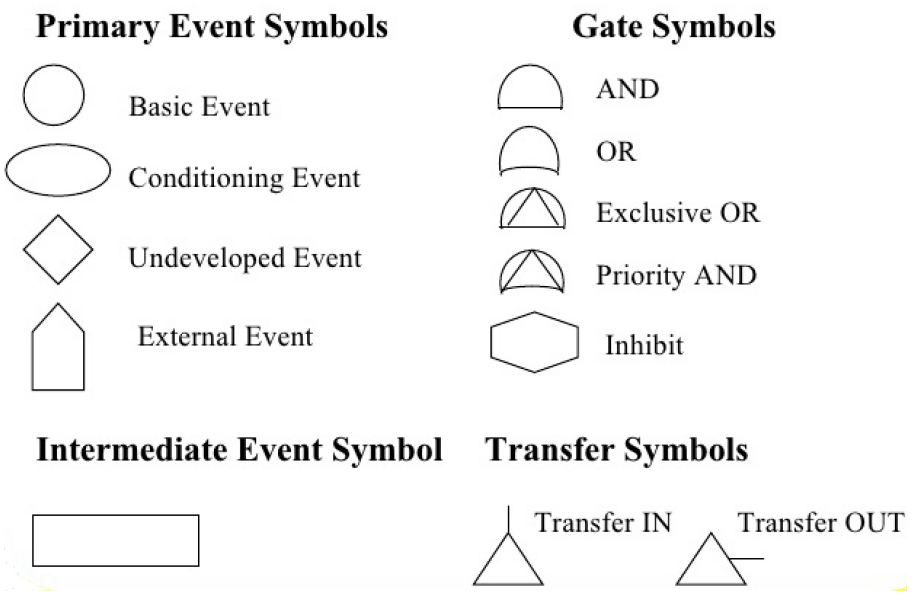


Fig. 2.1. Symbols in FTs

the designer an intuitive high-level abstraction of the system. Compared with the Failure Mode and Effects Analysis (FMEA), which is an inductive and bottom-up analysis method aimed at analyzing the effects of single component or function failures on equipment or subsystems, FTA is more useful in showing how resistant a system is to single or multiple initiating faults.

Fault Tree (FT) includes symbols (refer to Fig. 2.1) that show the basic events of the system, and the relation between these events and the state of the system [Ericson 1999]. The graphical symbols that show the relations are called logical gates. The output from a logical gate is determined by the input states. The system is analyzed in the context of its functional and safety requirements and environmental conditions. In FTs, the primary event symbols are given as follows:

- Basic event - failure or error in a system component or element (example: switch stuck in open position).
- Conditioning event - conditions that restrict or affect logical gates (example: mode of operation in effect).
- Undeveloped event - an event about which insufficient information is available, or which is of no consequence.
- External event - normally expected to occur (not of itself a fault).

An intermediate event gate can be used immediately above a primary event to provide more room to type the event description.

The gates work as follows:

- OR gate - the output occurs if any input occurs.
- AND gate - the output occurs only if all inputs occur (inputs are independent).
- Exclusive OR gate - the output occurs if exactly one input occurs.
- Priority AND gate - the output occurs if the inputs occur in a specific sequence specified by a conditioning event.
- Inhibit gate - the output occurs if the input occurs under an enabling condition specified by a conditioning event.

Transfer symbols are used to connect the inputs and outputs of related fault trees, such as the fault tree of a subsystem to its system.

In FTs, all combinations of basic events leading to the top event (TE) are identified. For example, as shown in Fig. 2.2, the TE is "D Fails". The basic events are linked to the TE, through intermediate events, by logical gates. A basic event does not necessarily represent a pure component failure. The basic events may include items such as hardware, various sub-systems, environmental factors, human error or some social matters. A standard FTA involves the following steps [Mahboob 2013]:

(1) Understand the system design and operation through data, drawings, procedures, diagrams, and so on.

(2) Define the problem and establish the correct TE (undesired events) for the analysis.

(3) Define the system rules and boundaries. What is included and what cannot be included?

(4) Follow the rules, boundaries, and logic (OR, AND,...) to build the FT model.

(5) Generate cut sets and compute probability values for the cut sets.

(6) Identify weak links and safety problems in the design and operation.

(7) Validate the FT model: check if the FT model is correct, complete, and accurately reflects system design and operation. Modify the FT if necessary during validation.

(8) Document the entire analysis with supporting data.

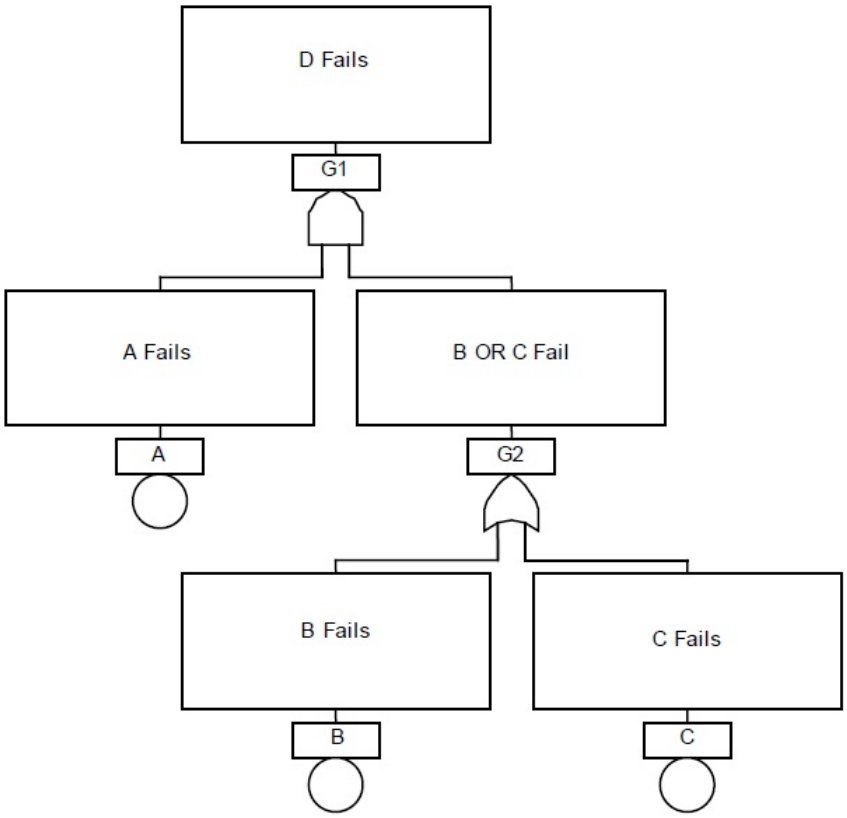


Fig. 2.2. An example of FT

In FTA, the concept of cut set is a group of basic events whose combined occurrence can cause the occurrence of the TE. A cut set will be minimal if it cannot be reduced further and still promises the occurrence of the TE. Each minimal cut set is considered as a parallel system of its components and the overall system state is considered as a series system of the minimal cut sets. The basic assumptions of the standard FTA include (1) the events in FTA represent random variables with binary states (occurring/not occurring) and (2) basic events are statistically independent. In general, the probability of TE in the FT is computed as a function of the minimal cut sets by using the inclusion and exclusion principle in Eq. (2.18) [Mahboob 2013].

$$P(TE) = \sum_{i=1}^{n} P(C_i) - \sum_{i=2}^{n} \sum_{j}^{i-1} P(C_i \cap C_j) + \cdots + (-1)^{n-1} P(C_1 \cap C_2 \cdots \cap C_n) \quad (2.18)$$

where $P(C_i)$ denotes the probability of the occurrence of minimal cut sets i in an FT and n is the number of minimal cut sets. The total number of events in a cut set is called the order of the cut set.

However, one obvious disadvantage of FTA is that it is not clear on failure mechanism,

since the causal relationship between events is not a simple YES or NO (1 or 0). Therefore, FTA is prone to missing the possible initiating faults. In addition, traditional static fault trees cannot handle the sequential interaction and functional dependencies between components. Consequently, it is necessary to employ dynamic methodologies to overcome these weaknesses.

2.3.2 Event Tree Analysis

Event Tree Analysis (ETA) is a forward and bottom-up logical modeling technique for both success and failure that explores responses through a single initiating event and lays a path for assessing probabilities of the outcomes and overall system analysis [Ericson 2015]. This analysis technique is first introduced during the WASH-1400 nuclear power plant safety study and widely used to analyze the effects of functioning or failed systems given that an event has occurred. ETA is a tool that will identify all consequences of a system that have a probability of occurring after an initiating event that can be applied to a wide range of systems including: nuclear power plants, spacecrafts, and chemical plants. This Technique may be applied to a system early in the design process to identify potential issues that may arise rather than correcting the issues after they have occurred. With this forward logic process, use of ETA as a tool in risk assessment can help to prevent negative outcomes from occurring by providing a risk assessor with the probability of occurrence. ETA uses a type of modeling technique called event tree (ET), which branches events from one single event using Boolean logic. As shown in Fig. 2.3, the ET begins with the initiating event where the consequences of this event follow in a binary (success/failure) manner. Each event creates a path in which a series of successes or failures will occur where the overall probability of occurrence for that path can be calculated. This process continues until the end state is reached. When the ET has reached the end state for all pathways, the outcome probability of the end state can be obtained.

Steps to perform an event tree analysis are shown as follows [Ericson 2015]:

(1) Define the system: Define what needs to be involved or where to draw the boundaries.

(2) Identify the accident scenarios: Perform a system assessment to find hazards or accident scenarios within the system design.

(3) Identify the initiating events: Use a hazard analysis to define initiating events.

(4) Identify intermediate events: Identify countermeasures associated with the specific scenario.

(5) Build the event tree diagram

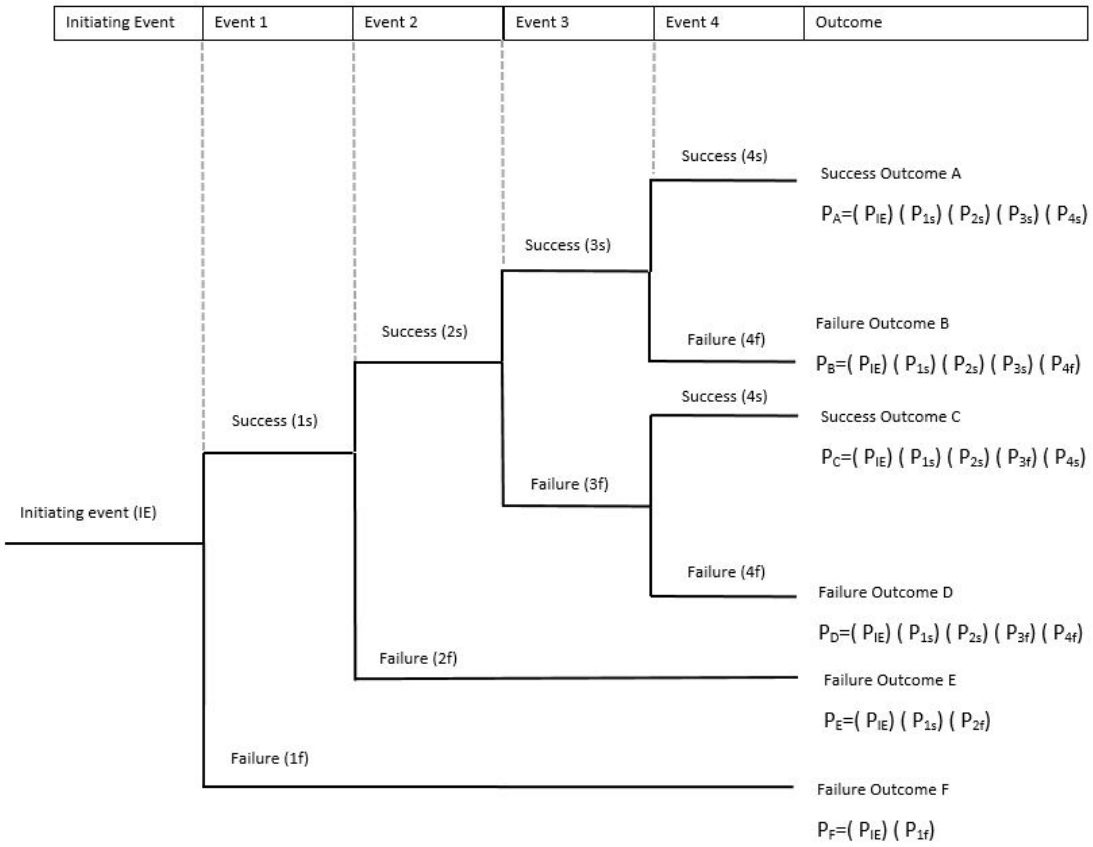


Fig. 2.3. An example of ET

(6) Obtain event failure probabilities: If the failure probability can not be obtained, use fault tree analysis to calculate it.

(7) Identify the outcome risk: Calculate the overall probability of the event paths and determine the risk.

(8) Evaluate the outcome risk: Evaluate the risk of each path and determine its acceptability.

(9) Recommend corrective action: If the outcome risk of a path is not acceptable develop design changes that change the risk.

(10) Document the ETA: Document the entire process on the event tree diagrams and update for new information as needed.

Obviously, ETA shows the following limitations: (1) it can address only one initiating event at a time. (2) Pathways must be identified in advance by the analyst. (3) Partial successes/failures are not distinguishable. Moreover, since the causal relationship between

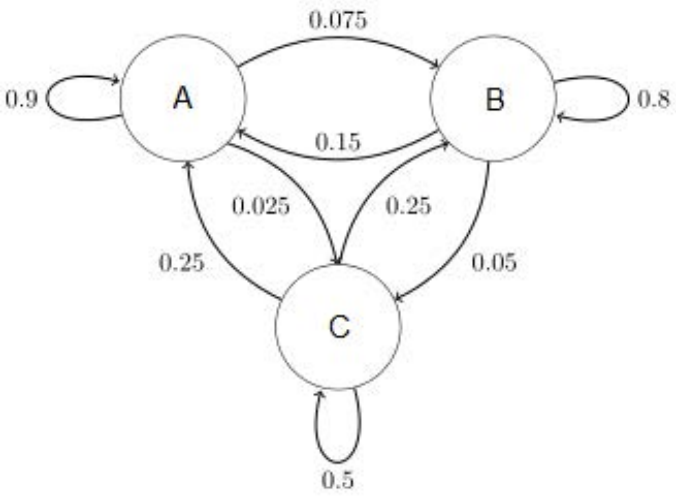


Fig. 2.4. An example of MC

events is not a simple "success" or "failure" in reality, it cannot preform thorough risk reasoning when it comes to deal with a complex system or scenario.

2.3.3 Markov Chains

Markov Chains (MCs) are stochastic models describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event and computing the probability that the system is in a specific state at a given time [Gagniuc 2017].

A MC is a type of Markov process that has either discrete state space or discrete index set (often representing time), but the precise definition of a MC varies. For example, it is common to define a MC as a Markov process in either discrete or continuous time with a countable state space (thus regardless of the nature of time), but it is also common to define a MC as having discrete time in either countable or continuous state space (thus regardless of the state space).

Fig. 2.4 shows an example of MC, which uses a directed graph to describe the state transitions.

The transition matrix for this example is expressed as follows:

$$P = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{array} \right] \end{array} \tag{2.19}$$

The distribution over states can be written as a stochastic row vector x with the relation

$x^{(n+1)} = x^{(n)}P$. So if at time n the system is in state $x^{(n)}$, then k time periods later, at time $n+k$ the distribution is calculated as follows:

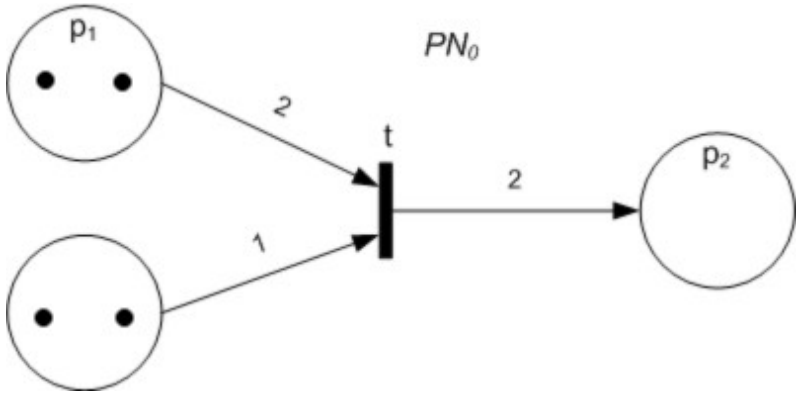
$$x^{(n+k)} = x^{(n+k-1)}P = (x^{(n+k-2)}P)P = \cdots = x^{(n)}P^k \tag{2.20}$$

MCs and their extensions have been mainly used for modeling complex dynamic system behavior and dependability analysis of dynamic systems. As an example, two-state Markov switching multinomial logit models are introduced by [Malyshkina 2009] to explain unpredictable, unidentified or unobservable risk factors in road safety analysis. Although MCs can elaborate the statistical state transition of different variables, they cannot formalize causal relationships between various events.

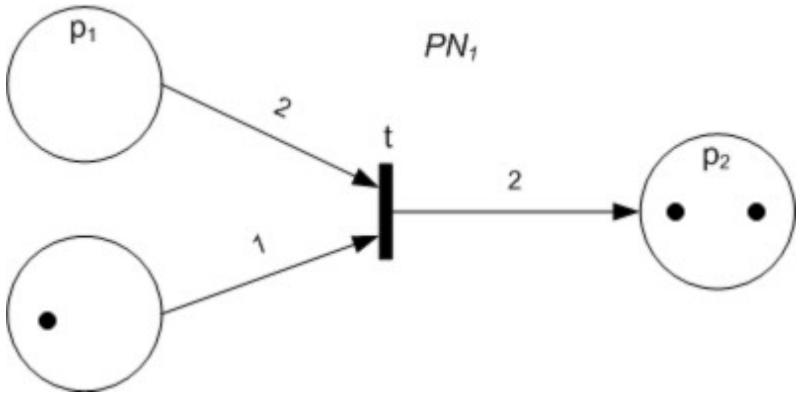
2.3.4 Petri nets

A Petri net (PN), also known as a place/transition (PT) net, is a graphical tool for the description and analysis of concurrent processes which arise in systems with several components (distributed systems). As shown in Fig. 2.5, a PN is a directed bipartite graph, which consists of places (i.e., conditions, represented by circles), transitions (i.e., events that may occur, represented by bars), and directed arcs. Arcs run from a place to a transition or vice versa, never between places nor between transitions. The places from which an arc runs to a given transition are called the input places of the transition; the places to which arcs run from a given transition are called the output places of the transition [Peterson 1981]. Places in a PN may contain a discrete number of marks called tokens. Any distribution of tokens over the places will represent a configuration of the net called a marking. In an abstract sense relating to a PN diagram, a transition of a PN may fire if it is enabled, i.e., there are sufficient tokens in all of its input places; when the transition fires, it consumes the required input tokens, and creates tokens in its output places. A firing is atomic, namely, a single non-interruptible step. Fig. 2.5 shows an example of PN.

As shown in Fig. 2.5a, the place p_1 is an input place of transition t; whereas, the place p_2 is an output place to the same transition. The configuration of PN_0 enables transition t through the property that all its input places have sufficient number of tokens (shown in the figures as dots) "equal to or greater than" the weights on their respective arcs to t. A transition can fire only if it is enabled. In this example, the firing of transition t results in configuration PN_1 as shown in Fig. 2.5b. In the diagram, the firing rule for a transition can be characterized by subtracting a number of tokens from its input places equal to the weight of their respective input arcs and accumulating a new number of tokens at the output places equal to the weight of the respective output arcs.



(a)



(b)

Fig. 2.5. An example of PN

A PN can also be defined as a net of the form $PN = (N, M, W)$, which extends the elementary net so that:

(1) $N = (P, T, F)$ is a net.

(2) $M : P \to Z$ is a place multiset, where Z is a countable set. M extends the concept of configuration and is commonly described with reference to PN diagrams as a marking.

(3) $W : F \to Z$ is an arc multiset, so that the count (or weight) for each arc is a measure of the arc weight.

Basic PNs have various extensions to integrate further features (time, probabilities, labeling, etc.), for instance, Colored Petri Nets (CPN) and Generalized Stochastic Petri Nets (GSPNs), etc. The selection of the PN variant depends on the features one wished to decide and the type of analysis to be carried out. For instance, in order to compare

the effectiveness of two main Automatic Protection Systems (APSs) at LXs, namely two-half-barrier APS and four-half-barrier APS, GSPNs were used in [Ghazel 2014] to analyze the aleatory fluctuations of various parameters involved in the dynamics within the LX area. However, the PN formalism does not enable, in a direct way, are unable to identify causality effectively when performing risk reasoning.

2.3.5 Fuzzy Logic

Fuzzy logic (FL) was introduced with the fuzzy set theory by Lotfi Zadeh [Cox 1994]. FL is a form of many-valued logic in which the truth values of variables may be any real number between 0 and 1. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false. By contrast, in Boolean logic, the truth values of variables may only be the integer values 0 or 1. Fuzzification operations can map mathematical input values into fuzzy membership functions. And the opposite de-fuzzifying operations can be used to map a fuzzy output membership functions into a "crisp" output value that can be then used for decision or control purposes. The process of Fuzzification is shown as follows:

(1) Fuzzify all input values into fuzzy membership functions.

(2) Execute all applicable rules (e.g., IF-THEN rules) in the rulebase to compute the fuzzy output functions.

(3) De-fuzzify the fuzzy output functions to get continuous output values.

FL works with membership values in a way that mimics Boolean logic. Namely, replacements for basic operators AND, OR, NOT must be available. There are several ways to reach this point. A common replacement is called the Zadeh operators as shown in Table 2.2:

Table 2.2. Zadeh operators

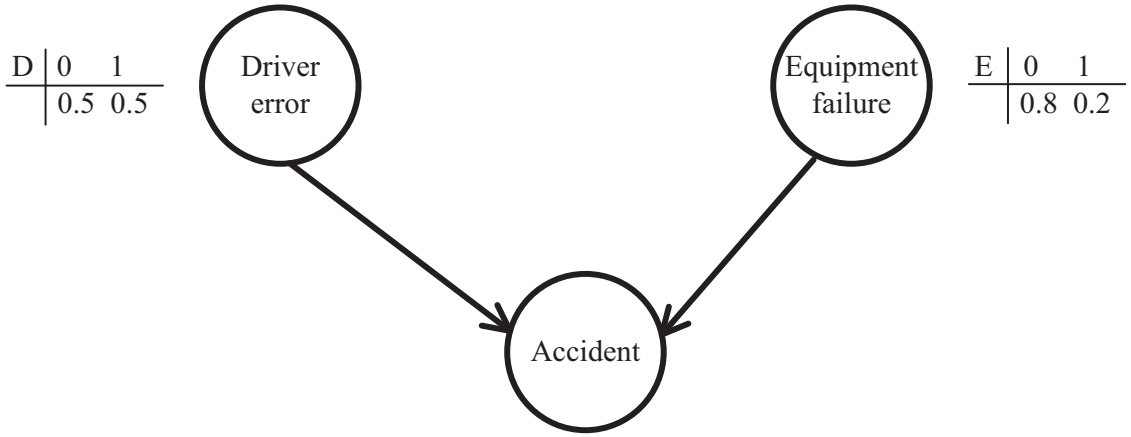
Boolean	Fuzzy
AND(x,y)	MIN(x,y)
OR(x,y)	MAX(x,y)
NOT(x)	$1 - x$

FL has been applied to many fields, from control theory to artificial intelligence. [Niittymaki 1998] discussed a FL controller for controlling the timing of a pedestrian crossing signal. The controller was designed to emulate the decision process of an experienced

crossing guard. However, the FL techniques are still unable to identify causality effectively when performing risk reasoning.

2.3.6 Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models that represent a set of variables and their conditional dependencies via a directed acyclic graph (DAG). They consist of a set of nodes which correspond to discrete or continuous random variables in Bayesian Networks and a set of directed links (arrows) which represent probabilistic dependence structure among nodes in Bayesian Networks. If the BNs consist of discrete random variables then each node will have a set of mutually exclusive and collective exhaustive states. A conditional probability table (CPT) is utilized to describe the conditional probability mass function of a discrete random variable. For example, in the BN shown in Fig. 2.6, "Driver error" (D) and "Equipment failure" (E) are parent nodes of "Accident" (A) and their CPTs are given.



D	0	1
	0.5	0.5

E	0	1
	0.8	0.2

D, E	Pr(A=0)	Pr(A=1)
0, 0	0.9	0.1
0, 1	0.1	0.9
1, 0	0.1	0.9
1, 1	0.05	0.95

Fig. 2.6. An example of BN

The joint probability function can be expressed as follows:

$$P(D, E, A) = P(D)P(E|D)P(A|D, E) \tag{2.21}$$

One can calculate $P(A=1)$ through forward inference:

$$\begin{aligned} P(A=1) = & P(A=1|D=0,E=0)P(D=0,E=0)+ \\ & P(A=1|D=1,E=0)P(D=1,E=0)+ \\ & P(A=1|D=0,E=1)P(D=0,E=1)+ \\ & P(A=1|D=1,E=1)P(D=1,E=1) \\ & = 0.585 \end{aligned} \tag{2.22}$$

On the other hand, one can calculate $P(D=1|A=1)$ through reverse inference:

$$\begin{aligned} P(D=1|A=1) &= \frac{P(D=1,A=1)}{P(A=1)} \\ &= \frac{P(D=1,A=1|E=0)P(E=0)+P(D=1,A=1|E=1)P(E=1)}{P(A=1)} \\ &= \frac{P(D=1,A=1,E=0)+P(D=1,A=1,E=1)}{P(A=1)} \\ &= 0.778 \end{aligned} \tag{2.23}$$

Over the last decade, BNs have been an increasingly popular notation used for risk analysis of safety-critical systems or large and complex dynamic systems [Chemweno 2015]. In order to obtain proper and effective risk control, risk planning should be performed based on risk causality, which can provide detailed information for decision making. In this context, a method for software risk analysis based on BNs combined with expert knowledge and V-structure discovery algorithm was proposed in [Hu 2010]. In [Lauria 2006], authors discussed how to build BNs from real-world data and incorporated BNs into decision support systems to support "what-if" analysis about Information Technology implementations. Heuristic model searching techniques and Maximum a Posteriori (MAP) estimation are used in this study to estimate the structure and the parameters of the BN. A semi-formal method for constructing the graphical structure of BNs based on domain knowledge using the causal mapping approach is discussed in [Nadkarni 2004]. The causal knowledge of experts is formally represented by causal maps, so as to consider the reasoning underlying the cause-effect relations perceived by individuals. In [Bouillaut 2013], Bouillaut et al. discussed the development of a decision tool realized by hierarchical Dynamic BNs (DBNs), which is dedicated to the maintenance of metro lines in Paris. This modeling work has comprehensively described the rail degradation process, the different diagnosis actors (devices and staff) and the decisions pertaining to maintenance actions. In [Langseth 2007], Langseth and Portinal discussed the applicability of BNs for reliability analysis and offered an instance of BNs' application for preventive maintenance. Moreover, the authors of this paper discussed the advantages behind BNs as follows: a) BNs constitute a modeling framework, which is particularly easy to use for interaction with domain experts; b) the sound mathematical formulation has been utilized in BNs to generate efficient learning methods; and c) BNs are equipped with an efficient calculation

scheme which often makes BNs preferable to traditional tools like Fault Trees (FTs). For more detailed knowledge about the application of BNs to risk assessment and decision support please refer to [Bensi 2010, Heredia-Zavoni 2012, Špačková 2013].

To sum up, the BN technique offers interesting features: flexibility of modeling, strong modeling power, high computational efficiency and, most importantly, the outstanding advantages involving the conjunction of domain expertise and automatic structure/parameters learning, causality analysis based on both forward inference (deductive reasoning) and reverse inference (abductive reasoning) [Weber 2012], as well as further influence and sensitivity analysis. For our study, given all the interesting features characterizing BNs, we adopt this notation for risk modeling, as discussed in chapter 6.

2.4 Summary

This chapter offers an overview on the LX safety analysis, statistical accident frequency prediction models and risk modeling techniques. Based on the investigation, we have identified a strong need for: (1) thorough quantitative analysis on motorist behavior with regard to various phases of the LX closure cycle, (2) an appropriate accident prediction model that should be comprehensive in terms of contributing factors to LX safety and (3) using BN technique to perform effective risk reasoning, so as to predict the accident occurrence and corresponding consequences and make cause diagnosis.

Part II

PRESENT WORKS AND CONTRIBUTIONS

Chapter 3

Preliminary Statistical Analysis on LXs

Sommaire

Overview

In this chapter, we focus on statistical risk analysis at SAL2 LXs. Various kinds of impacting factors, namely, transport mode, geographical region and traffic moment, are analyzed by means of statistical techniques to dig out their statistic and distribution characteristics based on the accident data from SNCF, the French railway operator. Then, we assess the effect of various factors on the risk level quantitatively, in such a way as to open the way for setting efficient solutions and consequently, reaching the point of improving LX safety.

The work reported in this chapter corresponds to the publication on the international conference "RAILWAYS 2016" [Liang 2016].

This chapter is structured as follows: Section 3.1 exposes a general introduction to LX safety statistics and potential impacting factors. Then, the general risk analysis of average accident frequency in terms of transport mode and geographical region is discussed in section 3.2. In section 3.3, the risk analysis related to frequency coefficient, is performed with regard to various traffic moment groups. Besides, the frequency coefficient distributed in different regions is further generated. Finally, section 3.4 concludes the main results and contributions of the present study.

3.1 Introduction

LX safety involves various aspects: technical elements, operational procedures, human factors and environmental considerations. On the other hand, the efficiency of several experiments adopted in the past years to improve LX safety (such as using obstacle detection at LXs, setting islets in front of LXs, etc.) is disputable due to the lack of thorough risk analysis establishing the potential relationship between impacting factors and the safety level. In [Jia 2007], impacts of different daily periods, seasons and railway lines on the probability of accidents occurring at LXs were analyzed, but some kinds of important factors, such as traffic moment, transport mode and geographical region are not taken into account. In [McCollister 2007], daily periods, train speeds and environmental factors at American LXs were analyzed and a statistical model was proposed to predict the probability of accidents. However, the predicted results have certain deviation with the actual situation due to the lack of thorough statistical analysis of the potential relationship between important impacting factors and accidents. In 2012, there were more than 118000 LXs in the 28 European countries, and 5 LXs per 10 line-km on average in Europe [ERA 2014]. In 2016, 111 train/vehicle collisions at French LXs led to 31 deaths [Plesse 2017]. Moreover, according to the statistics of SNCF, traffic moment, transport mode and geographical region are essential factors which shall be considered when improving the LX safety, nevertheless, these factors were not analyzed by the previous works, thoroughly.

Therefore, the objective of our study in this chapter is to make thorough analysis on various kinds of transport mode, different geographical regions and traffic moment to explore their influences on LX accidents. It should be noticed that the database from SNCF used in the analysis reported in this chapter contains detailed information about LX accidents/incidents from 1974 to 2014. Thereby, our analysis will consider the accident data during this period.

3.2 General risk analysis in terms of transport mode and geographical region

In this study, we considered the 21 geographical administrative regions in the mainland France divided in 2014. Accidents which are caused by the following main types of transport mode: 1) motorized vehicle (MV), 2) pedestrian or bicycle (PB), are considered respectively to allow for making statistical analysis in 21 different regions. The motorized vehicle contains the following sub-types:

- Car
- Bus
- Motorcycles with engine
- Truck
- Agricultural vehicle
- Exceptional convoy

The number of train-MV collisions and train-PB collisions, and the number of SAL2 in 21 regions are presented respectively in Fig. 3.1.

The general average frequency of accident occurrence per SAL2 per year are used to represent the general risk level involving total accidents, MV accidents, and PB accidents in different regions during the period considered. We can calculate the general average frequency through the Eq. (3.1), (3.2) and (3.3):

$$F_{G\_total_i} = \frac{Nb\_acc\_tot_i}{Nb\_SAL2_i \times Nb\_year}, i = 1, 2, \ldots, 21; \tag{3.1}$$

$$F_{G\_MV_i} = \frac{Nb\_acc\_MV_i}{Nb\_SAL2_i \times Nb\_year}, i = 1, 2, \ldots, 21; \tag{3.2}$$

$$F_{G\_PB_i} = \frac{Nb\_acc\_PB_i}{Nb\_SAL2_i \times Nb\_year}, i = 1, 2, \ldots, 21; \tag{3.3}$$

In Eq. (3.1), $F_{G\_total_i}$ represents the general average frequency of total accidents occurring in i^{th} region, $Nb\_acc\_tot_i$ represents the number of total accidents during a period of given years in i^{th} region, Nb_SAL2_i represents the number of SAL2 LXs in i^{th} region, and Nb_year represents the number of years of the period considered. In Eq. (3.2) and (3.3), $F_{G\_MV_i}$ represents the general average frequency of accidents caused by motorized vehicles occurring in i^{th} region, $Nb\_acc\_MV_i$ represents the number of accidents caused

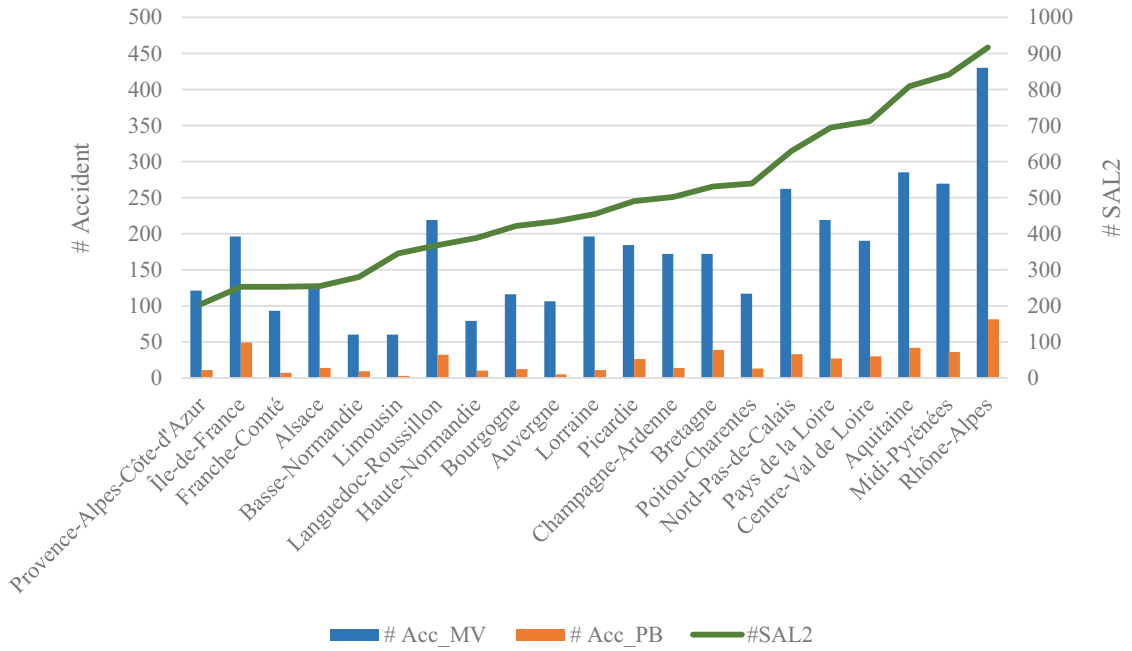


Fig. 3.1. Accidents caused by different transport modes at SAL2 in 21 French regions from 1974 to 2014

by motorized vehicles in i^{th} region, $F_{G\_PB_i}$ represents the general average frequency of accidents caused by pedestrians or bicycles occurring in i^{th} region, and $Nb\_acc\_PB_i$ represents the number of accidents caused by pedestrians or bicycles in i^{th} region.

Now that these three kinds of general average frequency in each region are determined, maps of French regions with the average accident frequency labeled are generated to show the risk distribution in different regions. As shown in Fig. 3.2, the general average frequency value of total accidents in the red region (greater than 0.02) is the highest. The general average frequency value of total accidents in the orange region (between 0.02 and 0.01) is at the medium level, and the general average frequency values of total accidents in the green region (less than 0.01) is the lowest. When we analyze the frequency figures in detail, we find that the risk is most serious in Île-de-France with a frequency of more than 0.02; Languedoc-Roussillon takes the second place with the frequency of about 0.017 followed by Provence-Alpes-Côte-d'Azur with the frequency of about 0.016. On the other hand, Limousin has the lowest risk with the frequency of about 0.005. Haute-Normandie and Basse-Normandie occupy the second and the third places of lowest risk successively. Turning to Fig. 3.1, the number of accidents occurring in Île-de-France is not the highest, but, when considering the corresponding number of LXs, this leads to the most serious risk. Conversely, although accidents in Rhône-Alpes are the most with the number of

521, the risk is not the highest due to a large number of LXs in this region. Limousin, Haute-Normandie and Basse-Normandie have significantly fewer accidents than any other region, less than 100 in the 40 years. That is why they have the lowest risk.

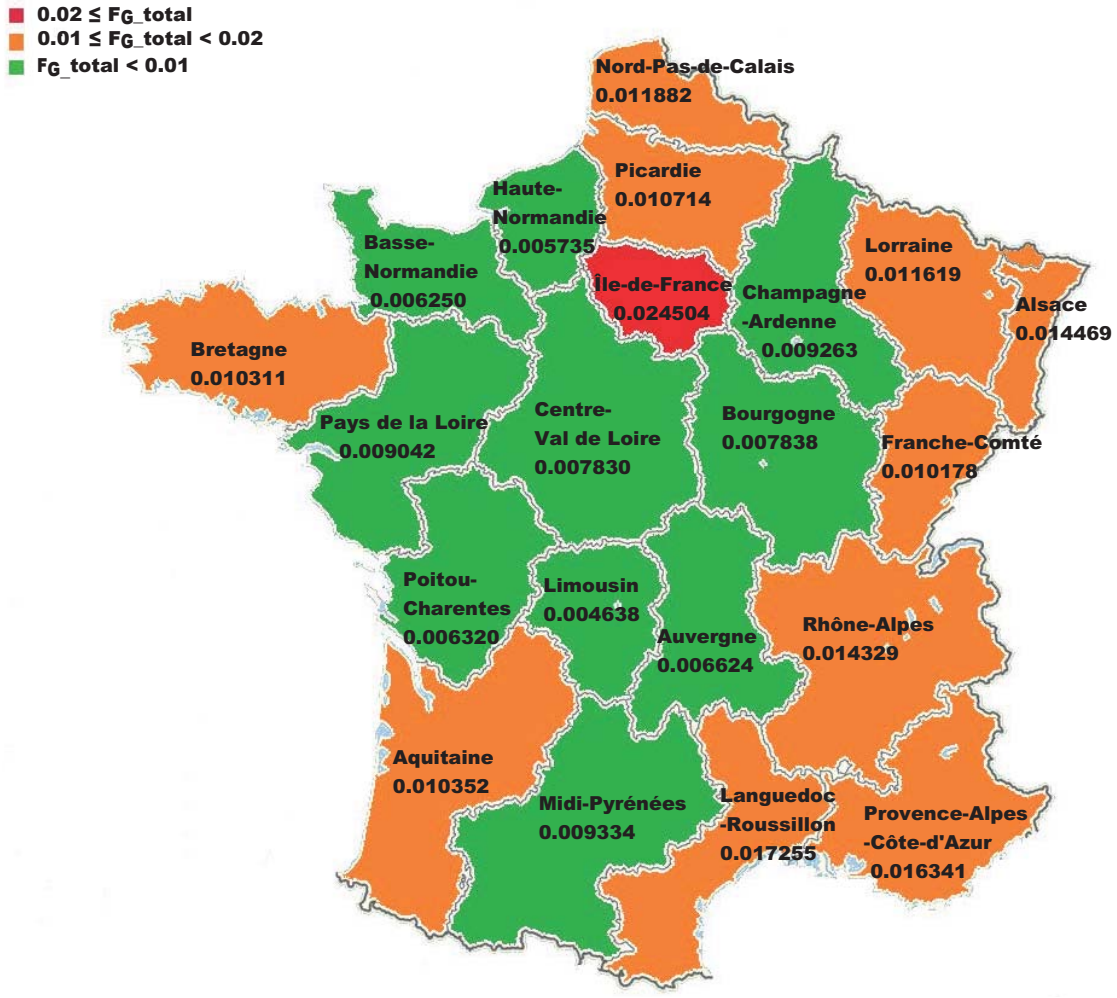


Fig. 3.2. General average frequency of total accidents distributing in different regions

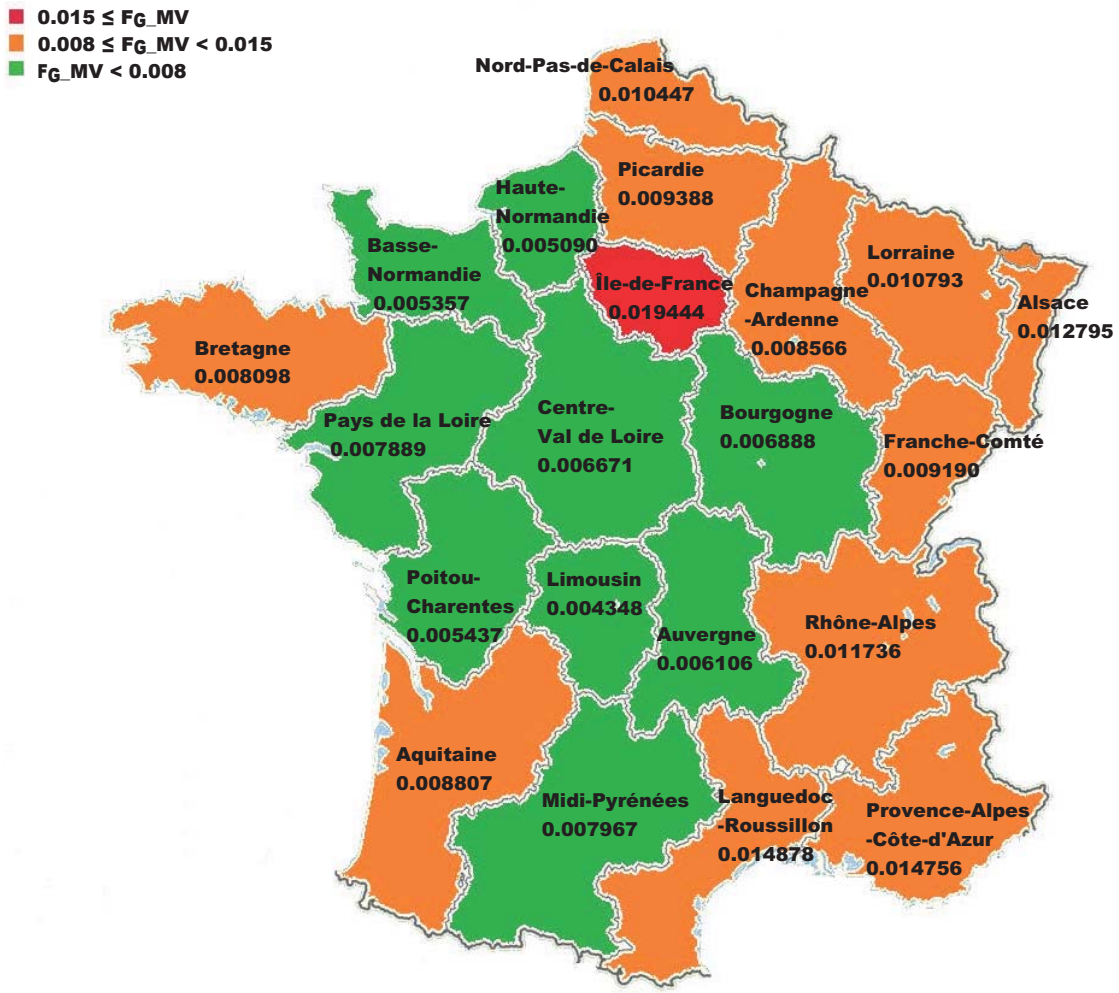


Fig. 3.3. General average frequency of MV accidents distributing in different regions

In Fig. 3.3 and Fig. 3.4, the general average frequency of accidents caused by motorized vehicles, pedestrians and bicycles in different regions is shown respectively. Considering the accidents caused by motorized vehicles, the distribution of frequency in different regions in Fig. 3.3 is relatively consistent with the distribution shown in Fig. 3.2. The only exception is Champagne-Ardenne. However, in Fig. 3.4, as for the accidents caused by pedestrians and bicycles, the distributions of frequency in different regions are relatively different from the distribution shown in Fig. 3.2.

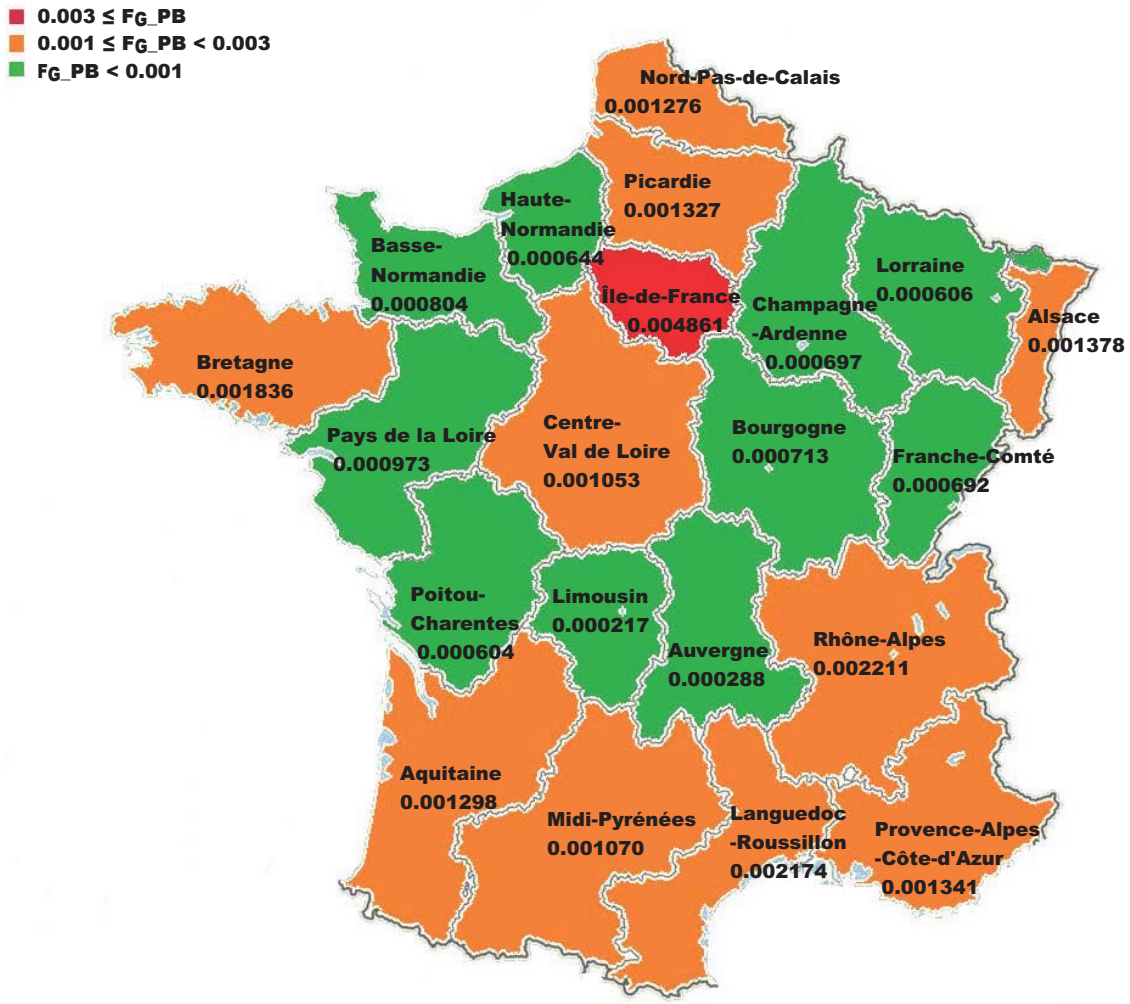


Fig. 3.4. General average frequency of PB accidents distributing in different regions

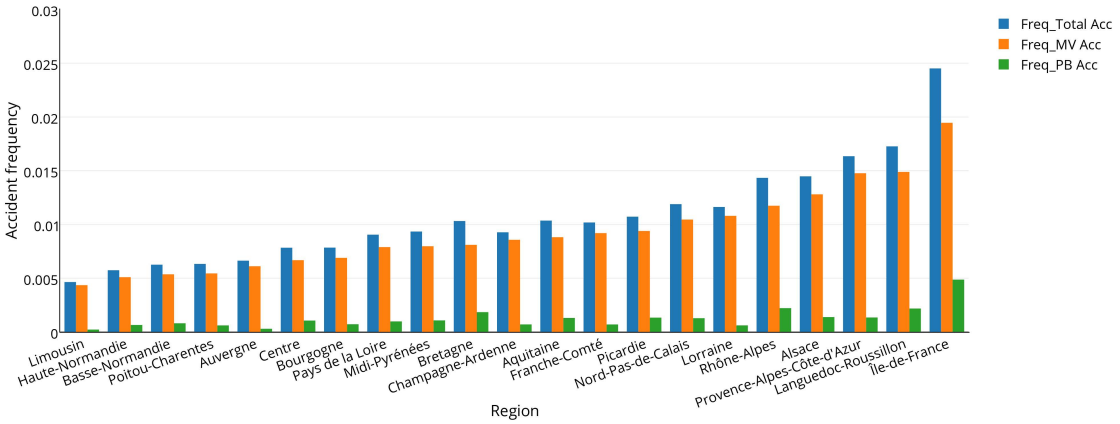


Fig. 3.5. Comprehensive general average frequency of accidents in different regions

According to the results shown in Fig. 3.2, 3.3, 3.4 and 3.5, we draw a conclusion that the main mode causing accidents at SAL2 is the motorized vehicle. Moreover, as the frequency of accidents involving motorized vehicle collisions increases, the total accident frequency in the overwhelming majority of regions increases accordingly, as illustrated in Fig. 3.5. On the contrary, pedestrians and bicycles contribute very little to the whole risk level and to the risk trend. Therefore, we can mainly focus on the solutions for reducing accidents caused by motorized vehicles to improve LX safety.

3.3 Risk analysis on frequency coefficient in terms of traffic moment

As mentioned in Section 3.2, the motorized vehicle is the main transport mode causing accidents at SAL2 LXs. Therefore, in this section we will focus on the impact of traffic moment on the whole risk level. We suspect that this parameter is one of the main parameters impacting LX risk level. Indeed, traffic moment gives the combined traffic (train/MV) at the LX and is defined as follows:

Definition

$$Moment = Road\ traffic\ frequency \times Railway\ traffic\ frequency \tag{3.4}$$

where $Road\ traffic\ frequency$ represents the average number of motorized vehicles per day at the LX, and $Railway\ traffic\ frequency$ represents the average number of trains per day crossing the LX. Here, we use "M" for short to denote the moment.

In this section, SAL2 LXs are classified by the category of M. In order to make the number of SAL2 in each M group to be as far as possible similar to any other group in each region, the moment groups have been defined in such a way that the number of SAL2 LXs in every group belongs to the interval [102, 155]. For example, in Table 3.1, there are three M categories in Auvergne, which are "$0 \leq M < 750$", "$750 \leq M < 5000$", "$5000 \leq M \leq 401850$", with the corresponding numbers of SAL2: 143, 144, 147, respectively. Besides, 401850 is the maximum M in this region. In this way, we can make risk analysis with regard to these SAL2 groups, thus making it possible to highlight the risk level related to different categories of M.

Table 3.1. Number of SAL2 in each region according to the category of "moment"

Number of SAL2 in each region according to the category of "moment"							
Auvergne	M [0, 750) 143	M [750, 5000) 144	M [5000, 401850] 147				
Haute-Normandie	M [0, 500) 133	M [500, 4500) 127	M [4500, 3082788] 128				
Midi-Pyrénées	M [0, 460) 142	M [460, 1150) 135	M [1150, 3000) 133	M [3000, 9000) 143	M [9000, 35000) 138	M [35000, 1169782] 141	
Basse-Normandie	M [0, 2100) 140	M [2100, 555600] 140					
Poitou-Charentes	M [0, 500) 133	M [500, 2200) 134	M [2200, 10500) 134	M [10500, 116978] 138			
Bourgogne	M [0, 1000) 142	M [1000, 8000) 141	M [8000, 720000] 138				
Limousin	M [0, 700) 115	M [700, 5000) 112	M [5000, 337524] 118				
Pays de la Loire	M [0, 500) 140	M [500, 1400) 137	M [1400, 3900) 143	M [3900, 15000) 138	M [15000, 815400] 136		
Champagne-Ardenne	M [0, 420) 122	M [420, 1800) 125	M [1800, 8000) 126	M [8000, 256410] 129			
Centre-Val de Loire	M [0, 350) 140	M [350, 1400) 143	M [1400, 4300) 144	M [4300, 18000) 141	M [18000, 688320] 144		
Aquitaine	M [0, 600) 135	M [600, 1600) 135	M [1600, 3800) 133	M [3800, 11000) 135	M [11000, 40000) 133	M [40000, 585644] 138	
Bretagne	M [0, 800) 131	M [800, 4200) 134	M [4200, 18000) 132	M [18000, 2244000] 134			
Rhône-Alpes	M [0, 320) 132	M [320, 1400) 128	M [1400, 3600) 131	M [3600, 11000) 131	M [11000, 35000) 132	M [35000, 100000) 133	M [100000, 1121088] 129
Franche-Comté	M [0, 2500) 125	M [2500, 271936] 128					
Picardie	M [0, 600) 121	M [600, 2550) 125	M [2550, 14000) 123	M [14000, 817000] 121			
Nord-Pas-de-Calais	M [0, 450) 124	M [450, 3000) 126	M [3000, 10000) 128	M [10000, 45000) 127	M [45000, 1477229] 124		
Lorraine	M [0, 1350) 155	M [1350, 8000) 152	M [8000, 402500] 150				
Provence-Alpes-Côte-d'Azur	M [0, 5700) 102	M [5700, 921250] 103					
Île-de-France	M [0, 11500) 127	M [11500, 752854] 126					
Alsace	M [0, 12000) 127	M [12000, 767088] 127					
Languedoc-Roussillon	M [0, 1250) 121	M [1250, 7200) 122	M [7200, 1428000] 125				

M [a, b) : a ≤ moment < b; M [c, d] : c ≤ moment ≤ d.

In order to analyze the impact of M, we firstly adopt the square root of M at SAL2 LXs to calculate a specific coefficient related to the average frequency of train-MV collisions (as a preliminarly involvement of the traffic moment factor), called frequency coefficient (FC), as shown in Eq. (3.5) (different from the general frequency statistics shown in Section 3.2). Logically, if the same number of accidents or more accidents occur during a given period at a SAL2 with small M than at another SAL2 with large M, this indicates that the SAL2 with small M has higher risk level than the SAL2 with large M. Furthermore, it is worth noticing that the SAL2 with small M accounts for a large proportion of the total SAL2

LXs in France. The formula adopted to establish the FC of collisions is the following:

$$FC_i = \frac{\sum_{j=1}^{Nb\_SAL2\_reg_i} \frac{Nb\_acc\_MV_j}{\sqrt{M_j}}}{Nb\_SAL2\_reg_i \times Nb\_year},\ i = 1, 2, \ldots, 21; \tag{3.5}$$

where FC_i represents the frequency coefficient of accidents caused by train-MV collisions in i^{th} region; $Nb\_acc\_MV_j$ represents the number of accidents caused by train-MV collisions at j^{th} SAL2 LX during a period of given years in i^{th} region; M_j represents the moment at the j^{th} SAL2 LX in i^{th} region; $Nb\_SAL2\_reg_i$ represents the number of SAL2 in i^{th} region, and Nb_year represents the number of years of the considered period.

The FC of collisions regarding the different M categories in each region are shown in Table 3.2. The M categories marked with orange color have the most serious risk in various regions. We find that SAL2 LXs in the category of smallest M have the highest risk level in 86% of the regions, except in Auvergne, Champagne-Ardenne and Nord-Pas-de-Calais where the second smallest M have the highest risk level. On the other hand, SAL2 LXs in the category of largest M have the lowest risk level in 81% of the regions. These statistical results powerfully prove the logic inference mentioned above that the SAL2 with small M has higher risk level than the SAL2 with large M when considering the ratio of the number of accidents occurring per year at the SAL2 to the moment value of the SAL2.

In order to further analyze the impact of M on the risk level in each region, the FC of train-MV collisions occurring at all the SAL2 LXs in each region is determined.

Table 3.2. Frequency coefficient of MV accidents at SAL2 in each region according to the category of "moment"

Frequency coefficient of MV accidents at SAL2 in each region according to the category of "moment"							
Auvergne	M[0, 750) 0.000049	M[750, 5,000) 0.000100	M[5,000, 401,850] 0.000078				
Haute-Normandie	M[0, 500) 0.000190	M[500, 4,500) 0.000089	M[4,500, 3,082,788] 0.000072				
Midi-Pyrénées	M[0, 460) 0.000202	M[460, 1,150) 0.000068	M[1,150, 3,000) 0.000072	M[3,000, 9,000) 0.000066	M[9,000, 35,000) 0.000095	M[35,000, 1169,782] 0.000065	
Basse-Normandie	M[0, 2,100) 0.000147	M[2,100, 555,600] 0.000067					
Poitou-Charentes	M[0, 500) 0.000302	M[500, 2,200) 0.000064	M[2,200, 10,500) 0.000098	M[10,500, 1169,782] 0.000078			
Bourgogne	M[0, 1,000) 0.000224	M[1,000, 8,000) 0.000130	M[8,000, 720,000] 0.000061				
Limousin	M[0, 700) 0.000269	M[700, 5,000) 0.000058	M[5,000, 337,524] 0.000055				
Pays de la Loire	M[0, 500) 0.000175	M[500, 1,400) 0.000121	M[1,400, 3,900) 0.000111	M[3,900, 15,000) 0.000122	M[15,000, 815,400] 0.000062		
Champagne-Ardenne	M[0, 420) 0.000129	M[420, 1,800) 0.000285	M[1,800, 8,000) 0.000099	M[8,000, 256,410] 0.000185			
Centre-Val de Loire	M[0, 350) 0.000201	M[350, 1,400) 0.000121	M[1,400, 4,300) 0.000092	M[4,300, 18,000) 0.000087	M[18,000, 688,320] 0.000059		
Aquitaine	M[0, 600) 0.000162	M[600, 1,600) 0.000135	M[1,600, 3,800) 0.000095	M[3,800, 11,000) 0.000093	M[11,000, 40,000) 0.000091	M[40,000, 585,644] 0.000055	
Bretagne	M[0, 800) 0.000324	M[800, 4,200) 0.000117	M[4,200, 18,000) 0.000068	M[18,000, 2244,000] 0.000044			
Rhône-Alpes	M[0, 320) 0.000297	M[320, 1,400) 0.000122	M[1,400, 3,600) 0.000106	M[3,600, 11,000) 0.000101	M[11,000, 35,000) 0.000080	M[35,000, 100,000) 0.000054	M[100,000, 1121,088] 0.000047
Franche-Comté	M[0, 2,500) 0.000143	M[2,500, 271,936] 0.000101					
Picardie	M[0, 600) 0.000314	M[600, 2,550) 0.000135	M[2,550, 14,000) 0.000127	M[14,000, 817,000] 0.000101			
Nord-Pas-de-Calais	M[0, 450) 0.000079	M[450, 3,000) 0.000203	M[3,000, 10,000) 0.000082	M[10,000, 45,000) 0.000063	M[45,000, 1477,229] 0.000059		
Lorraine	M[0, 1,350) 0.000250	M[1,350, 8,000) 0.000122	M[8,000, 402,500] 0.000099				
Provence-Alpes-Côte-d'Azur	M[0, 5,700) 0.000212	M[5,700, 921,250] 0.000094					
Île-de-France	M[0, 11,500) 0.000285	M[11,500, 752,854] 0.000099					
Alsace	M[0, 12,000) 0.000513	M[12,000, 767,088] 0.000113					
Languedoc-Roussillon	M[0, 1,250) 0.000442	M[1,250, 7,200) 0.000134	M[7,200, 1428,000] 0.000152				
M[a, b): a ≤ moment < b; M[c, d]: c ≤ moment ≤ d.							

Fig. 3.6 shows the FC acted by the square root of M distributed over the different regions. It shows that the risk level is the highest in Île-de-France region with the FC of 7.31×10^{-4}; Alsace region takes the second place followed by Languedoc-Roussillon and Provence-Alpes-Côte-d'Azur. Through detailed analysis, we find that the general average frequencies of MV accidents in these 4 regions (see Section 3.2) are also the highest. Moreover, according to the recorded statistics, more train-MV collisions happened at SAL2 with small M in Île-de-France than in the other 3 regions during the period considered.

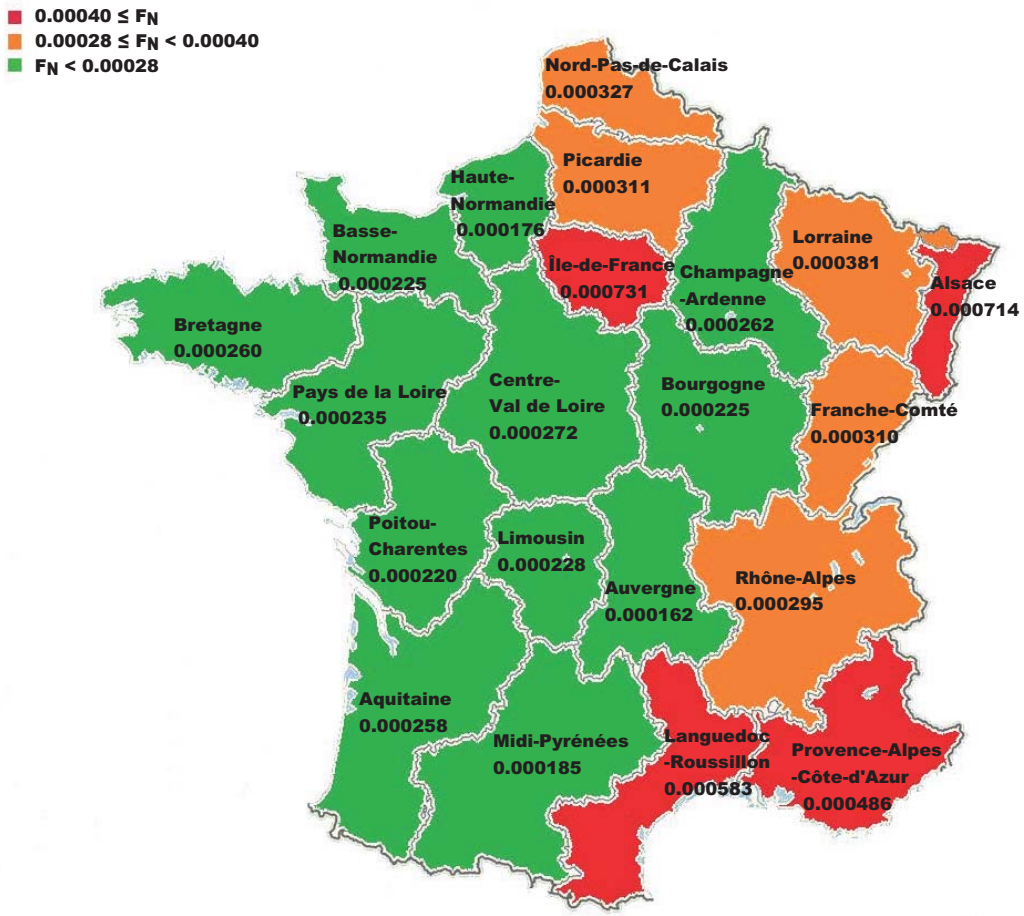


Fig. 3.6. Frequency coefficient of MV accidents distributing in different regions

3.4 Summary

This chapter presents risk analysis relative to LXs based on recorded accident statistics. Various parameters have been taken into account in our study: the involved road transport mode, geographical regions and the traffic moment. First, general risk analysis of the entire accidents, train-MV accidents and train-PB accidents in different French regions is performed. Then further investigation is carried out to analyze the risk of train-MV accidents in terms of frequency coefficient acted by the square root of "moment" according to various "moment" classification in the different regions. Based on the obtained results, we draw a conclusion that more attention shall be paid on the SAL2 LXs in 4 specific regions (Île-de-France, Languedoc-Roussillon, Alsace, and Provence-Alpes-Côte-d'Azur). In particular, at a micro level, for an individual SAL2 LX, the smaller the "moment" it has, the higher the risk it may show.

The thorough statistical analysis presented in this chapter allows us to identify the main risk factors and quantify their impacts on the overall LX risk. Although the analyses reported in this chapter are based on the accident data of France and focus on the SAL2 LXs, they have an important reference value when carrying out risk analysis at LXs in other countries. Moreover, the methodology can be easily adopted in other countries.

As reported in this chapter that motorized vehicle is main transport mode causing accidents at SAL2 LXs, we will focus on the studies related to train-MV accidents at SAL2 LXs in following chapters.

Chapter 4

Motorist Behavior Quantitative Analysis: Experiments at 12 selected automated LXs

Sommaire

Overview

In chapter 3, we noticed that violations committed by motorists are the primary cause of LX accidents. Therefore, vehicle driver behavior at LXs is a safety concern that requires a particular care, in order to set an appropriate diagnosis. The study in this chapter is a tentative to acquire a better understanding of risky vehicle driver behavior while crossing LXs during the closure cycle. A risk analysis of motorist behavior is performed based on field measurement conducted at 12 automatic LXs (11 equipped with two half barriers (SAL2) and 1 equipped with four half barriers (SAL4)). Thanks to suitable instruments set at the LX entrance from either side, various factors have been analyzed. Hence, we focus on motorist behavior during the LX closure cycle while distinguishing between different phases. Namely, the closure cycle is divided into three phases which are "Ph2 Red Flash and Siren", "Ph3 Barriers Coming Down" and "Ph4 Barriers Down". A statistical analysis is subsequently performed according to the phase periods. Furthermore, vehicle

driver behavior in each phase as time increases is scrutinized respectively. The relationship between vehicle driver behavior and specific ranges of vehicle speed is examined. In particular, motorist behavior during Ph4 is detected and analyzed in detail.

The work reported in this chapter corresponds to the publications on the journals "Accident Analysis & Prevention" [Liang 2017b] and "Safety Science" [Liang 2018c].

This chapter is structured as follows: section 4.1 exposes a general introduction to motorist behavior analysis for LX safety and the definition of LX closure cycle. Then, the motorist behavior analysis related to Ph2 and Ph3 is discussed in section 4.2. In section 4.3, the motorist behavior analysis related to Ph4 is performed to investigate the zigzag scenarios. Section 4.4 examines the distinction of motorist responses to SAL2 and SAL4 protection systems, so as to compare the efficiency of SAL2 and SAL4 LXs in terms of LX safety. A detailed discussion about the obtained results is given in section 4.5. Finally, section 4.6 offers a general summary of this chapter.

4.1 Introduction

As mentioned in chapter 2, although research on human factors related to LX safety has been an area of great concern over the past decades [Wilson 2005, Wilson 2014], the causes of collisions that occur at LXs remain insufficiently understood. It worth noticing that few existing works focus on the analysis of motorist behavior with regard to the separate phases of the automated LX closure cycle. Some available studies adopted qualitative approaches for understanding motorist behavior during the entire cycle of LX control, including when the LX is open to road traffic [Shappel 2000, San Kim 2013].

In general, the causes of such collisions fall into two broad classes, which are unintentional error and intentional violation. As for the unintentional errors, vehicle drivers may, for instance, fail to observe the warnings or fail to determine the braking distance appropriately. However, as for the intentional violations, vehicle drivers observe the warnings and fully understand their meaning, but intentionally commit transgressions on their own judgment [Lenné 2011]. In the present study, we carry out experiments that aim to quantitatively and finely analyze intentional violations of motorists during the LX closure cycle. These experiments have been conducted at 12 automated LXs in France, among which there are 11 LXs equipped with two half barriers (SAL2) (four SAL2 LXs are firstly selected for the analysis during Ph2 and Ph3; seven more SAL2 LXs are selected later for the analysis during Ph4) and 1 equipped with four half barriers (SAL4).

Here, we need to introduce that the control cycle of SAL2 LXs consists of five phases:

- Ph0 “Unflash and Barriers Coming Up”: corresponds to the phase launched right after the crossing train leaves the intersection zone. In this phase, the warning lights stop flashing, the sirens stop blaring and the barriers are rising.

- Ph1 “Barriers Up”: corresponds to the phase when the LX is open for road traffic as no train is approaching. In this phase, the barriers are fully opened.

- Ph2 "Red Flash and Siren": in this phase, the warning lights flash and sirens whistle, whereas the barriers are kept raised.
- Ph3 "Barriers Coming Down": in this phase, the barriers are falling.
- Ph4 "Barriers Down": in this phase, the barriers are fully lowered and the lights keep flashing.

Ph2, Ph3 and Ph4 are the three phases of the closure cycle of LX control.

In the sequel, based on recorded measurements, the zigzag violation rate is analyzed with regard to the hour, the weekday and the vehicle speed during Ph2 and Ph3. Moreover, the zigzag violation rate in Ph4 is further investigated with regard to the prolonged LX closure duration and LX location (railway station nearby or not), respectively. Then, some other features characterizing risky behavior are determined, such as troop phenomenon, etc. Besides, since our aim is to analyze motorist behavior at SAL2 LXs, one SAL4 LX was considered in our experiments to examine the distinction of motorist responses to SAL2 and SAL4 protection systems, so as to compare the efficiency of SAL2 and SAL4 LXs in terms of safety.

4.2 Motorist behavior analysis during Ph2 and Ph3

In this stage, four LXs, named Motteville (LX 51), Ectot lès Baons (LX 55), Yvetot (LX 58) and Gonfreville (LX 69), were selected firstly for our field measurement campaign related to Ph2 and Ph3. As shown in Fig. 4.1, the four red circular marks represent the accurate locations of the selected LXs and the corresponding kilometer points (KP) are indicated. The railway line traversing these four LXs is a bidirectional double-track line. In addition, the railway speed limit at these four LXs is 160 km/h. The railway stations in the vicinity of these four LXs are also indicated (see triangle marks in Fig. 4.1). LX 51 and LX 55 are on the right side and left side of Motteville railway station, respectively. It can also be seen that LX 51, LX 55 and LX 58 are close to each other, whereas LX 69 is about 40 kilometers from LX 58. The four LXs were selected based on various features (e.g., environment, infrastructure, equipment, railway and road traffic, etc.) as well as the accident/incident statistics (see Table 4.1). The LX selection process ensures that a range of LX features is represented including a variety of road traffic density (enabling analysis of the impact of road traffic volume) and a variety of LX closure durations. In particular, the choice of some LXs close to railway stations ensures long closure cycles in some cases. Moreover, we assume that such long closure cycles would potentially foster "zigzag" vehicle driver behavior. Statistics in terms of LX accidents/incidents have also

Fig. 4.1. Geographic information about LX 51, LX 55, LX 58 and LX 69

been considered to select the LXs for our measurement. For this aim, we have analyzed the LX accident/incident data from 1974 to 2014, which were collected by SNCF and recorded in a dedicated database. It is shown in Table 4.1 that 3 collisions and 114 near misses occurred at LX 55 during the period concerned, which take the first place among the four LXs considered. LX 51 with 3 collisions and 17 near misses occupies the second place. LX 58 and LX 69 had no collision during the last 40 years and only 5 near misses and 1 near misses, respectively.

Table 4.1. Collisions/near misses and road/rail traffic volumes at the four LXs from 1974 to 2014

LX	# Collisions	# Near misses	# Average daily road traffic	# Average daily train traffic
51 Motteville	3	17	9455	89
55 Ectot lès Baons	3	114	21491	70
58 Yvetot	0	5	13938	89
69 Gonfreville	0	1	10946	91

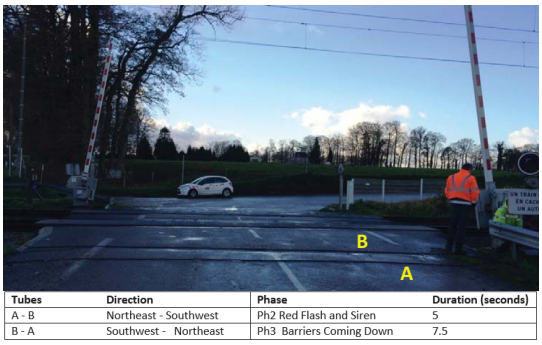
4.2.1 Road traffic measurement

This field observation campaign conducted by SNCF Réseau in cooperation with Metro-Count, an engineering office specializing in road traffic measurement. The sensing equipment installed at each LX consists of two separate tubes one meter apart. These two

tubes are placed perpendicularly to road traffic direction on the ground in front of the stop line, just before the level crossing barriers. They are connected to a counter that is a digital recording module, which, in turn, is linked to the light flashing signal to ensure synchronization with the signal cycle. The synchronization between the installed counting module and the LX control module is achieved by means of an optical sensing device installed in front of the LX flashing lights in both directions. Thanks to the installation set, the timestamp, vehicle direction, vehicle speed and phase, are recorded for each crossing vehicle individually.

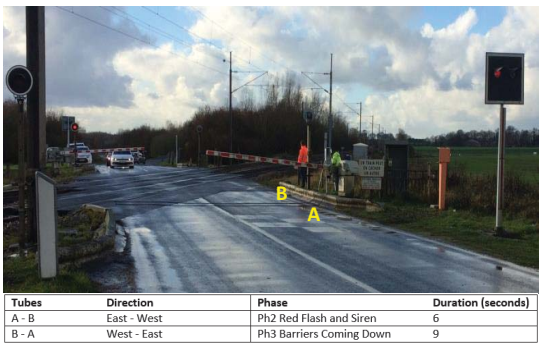
The measurement was performed over a period of 9 weeks. There were 86928, 192123, 129661 and 98122 vehicles in total crossing LX 51, 55, 58 and 69 during 5249, 5058, 2668 and 4255 LX control cycles respectively, in the measurement period of 9 weeks. Fig. 4.2 shows labeled images taken from the four LXs considered, as well as some related information. For instance, Fig. 4.2a shows the warning devices and measurement instruments installed at LX 51. One can notice two sensing tubes which are installed on the ground surface in front of the LX. The duration of Ph2 and Ph3 at LX 51 is 5s and 7.5s, respectively.

The field traffic data collected through our installation are recorded in a specific database. As shown in Fig. 4.3a, a tool, called "MTExec", is used to visualize and handle the recorded data. With the help of this tool, one can select different phases, as well as the specific beginning time and ending time of each phase, etc. This allows for different kinds of statistical analysis on various impacting factors. Fig. 4.3b shows an example of road traffic data that are recorded in the MTExec tool. As can be noted, the timestamp, vehicle direction, vehicle speed, vehicle type and phase information are recorded distinctly. Based on these records, the vehicle dynamics can be fully characterized.



Tubes	Direction	Phase	Duration (seconds)
A - B	Northeast - Southwest	Ph2 Red Flash and Siren	5
B - A	Southwest - Northeast	Ph3 Barriers Coming Down	7.5

(a) LX 51 Motteville



Tubes	Direction	Phase	Duration (seconds)
A - B	East - West	Ph2 Red Flash and Siren	6
B - A	West - East	Ph3 Barriers Coming Down	9

(b) LX 55 Ectot lès Baons

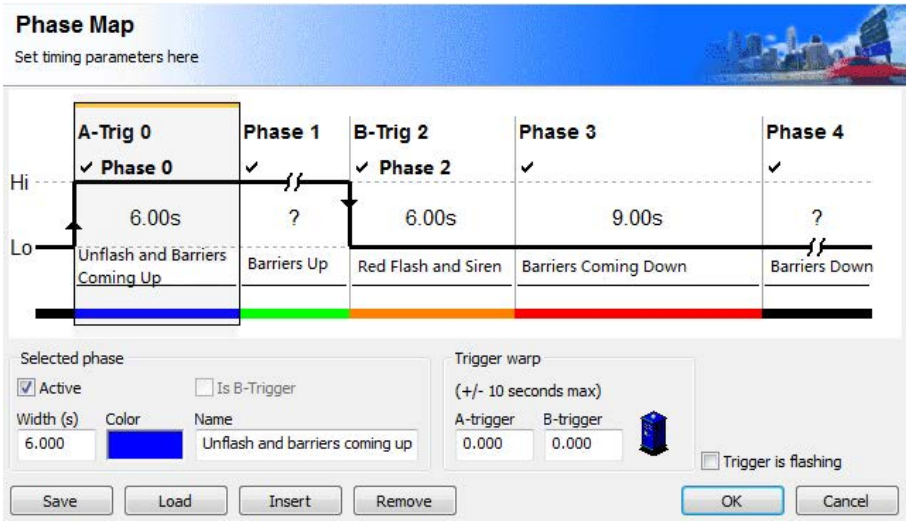
Tubes	Direction	Phase	Duration (seconds)
A - B	North - South	Ph2 Red Flash and Siren	8
B - A	South - North	Ph3 Barriers Coming Down	7

(c) LX 58 Yvetot

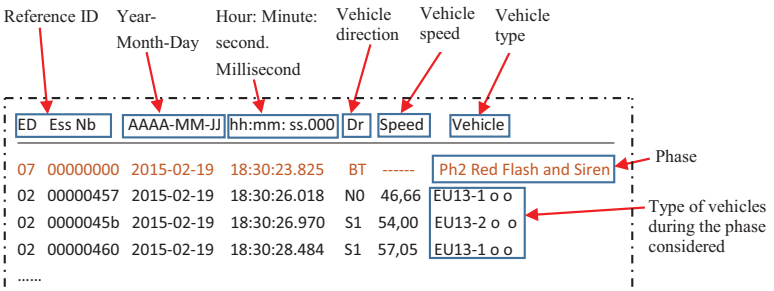
Tubes	Direction	Phase	Duration (seconds)
A - B	South - North	Ph2 Red Flash and Siren	7
B - A	North - South	Ph3 Barriers Coming Down	9

(d) LX 69 Gronfreyville

Fig. 4.2. Labeled photographs of environment and devices at the four LXs



(a) The tool "MTExec"



(b) Example of road traffic data recorded

Fig. 4.3. The tool "MTExec" and an example of road traffic data recorded in the MTExec tool

It should be noted that, in terms of ethics approval, the data collected during the measurement do not hold any personal or private aspects.

4.2.2 Behavioral analysis during Ph2 and Ph3

Motorist behavior during Ph2 and Ph3 is discussed in this section. We firstly investigate how the number of crossing vehicles evolves as time advances during Ph2 and Ph3 (section 4.2.2.1). Then, various parameters are taken into account for the analysis of the violation rate during Ph2 and Ph3, namely, time-slots of the day and weekday (section 4.2.2.2). The speed of violating vehicles is investigated as well later on in section 4.2.2.3.

4.2.2.1 Global violation trend during Ph2 and Ph3

In this section, we analyze how the number of violating vehicles evolves as time increases along Ph2 and Ph3. Thus, the observation time slot considered lasts from the beginning of Ph2 until the end of Ph3. Moreover, a distinction will be made between daytime and night. The daytime is from 6:00 h to 21:00 h and the night-time ranges from 21:00 h to 6:00 h.

When vehicle drivers observe the flashing lights and hear sirens, then, if it is possible to stop before the LX, they should brake. The whole process takes a period of reaction time. According to ergonomics studies, human visual reaction time is about $0.2 \sim 0.25$ seconds [Green 1967] and auditory reaction time is $0.12 \sim 0.18$ seconds [Taylor 1967]. Moreover, the human nerve transfer takes around 0.5 seconds of the refractory period in general [Welford 1952]. Thus, the whole operating delay of the action guided by sense organs should be less than 1 second. In addition, the longest braking coordination time of motorized vehicles does not exceed 0.8 seconds. Then, the braking delay itself should be considered. Therefore, a vehicle crossing during the first 4 seconds of Ph2 will not be regarded as a violation in our study, only a vehicle crossing subsequently to the first 4 seconds should be considered as a violation, which may correspond to either an intentional violation or the inability of the motorist to stop his vehicle safely due to a high speed. As mentioned above, the speed of violating vehicles will be investigated later on in section 4.2.2.3.

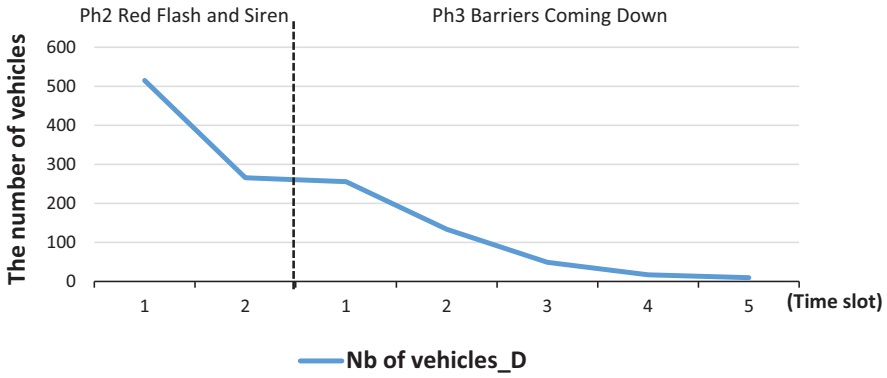
Turning to data as shown in the table of Fig. 4.2, 1, 2, 4 and 3 seconds are left after

OBS - Date	OBS - Hour	OBS-Week	RP	OBS - Number	OBS - Direction	OBS - Date of departure	OBS - Hour of departure
14/01/2015	4:51:29	Mercredi	YVETOT - BV	1421212074	Even	14/01/2015	1:52:00
14/01/2015	5:20:30	Mercredi	YVETOT - BV	1421212554	Odd	14/01/2015	4:36:00

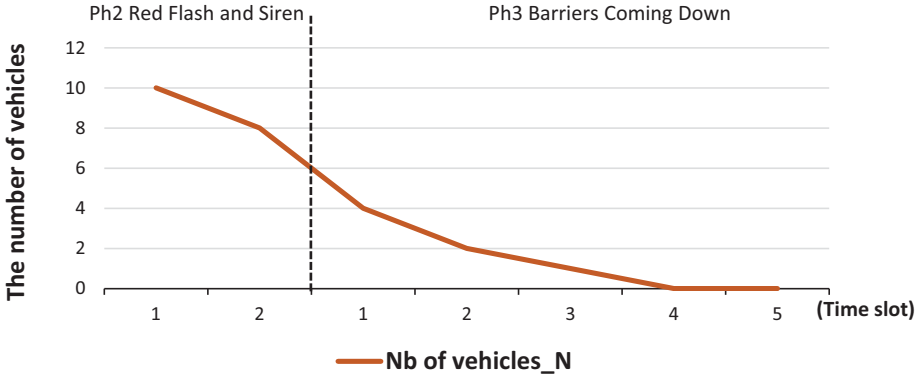
Fig. 4.4. An example of detailed railway traffic data offered by SNCF

excluding the first 4 seconds of Ph2 at LX 51, LX 55, LX 58 and LX 69, respectively. Indeed, the duration of Ph3 corresponding to these four LXs is 7.5, 9, 7 and 9 seconds, respectively. Due to the discrepancy in terms of duration considered, instead of considering the successive seconds in each phase (Ph2 and Ph3), we divide Ph2 (resp. Ph3) into 2 (resp. 5) equal time intervals at each LX and then, we can make the analysis in terms of phase halves: 1^{st} and 2^{nd} (resp. phase quintiles: 1^{st}, 2^{nd}, 3^{rd}, 4^{th} and 5^{th}) in such a way as to be able to merge the road traffic data recorded at the four LXs. As we seek to investigate the general trend of violation volume as time increases along Ph2 and Ph3, we will consider the number of violating vehicles rather than the corresponding rate.

As shown in Fig. 4.5a, "Nb of vehicles _ D" represents the number of vehicles crossing LX during Ph2 and Ph3 in the daytime. The profile declines dramatically as time advances during Ph2 and the violating vehicles during Ph3 are much fewer. This can be explained by the fact that vehicle drivers are aware of the increasing risk as time passes during the closure cycle. In Fig. 4.5b, "Nb of vehicles _ N" represents the number of violating vehicles during Ph2 and Ph3 at night. In fact, the vehicles crossing LXs at night are significantly fewer than those in the daytime. More importantly, the profile declines more slowly at night than during the daytime at the beginning of Ph2. A potential reason is that the visibility is worse at night; consequently, drivers need to take a longer time to watch the lights and the surroundings.



(a) Violation volume trend at the four LXs during Ph2 and Ph3 in the daytime



(b) Violation volume trend at the four LXs during Ph2 and Ph3 at night

Fig. 4.5. Violation trend at the four LXs as time increases in the daytime and at night during Ph2 and Ph3

4.2.2.2 Analysis of the violation rate according to the week and the hour

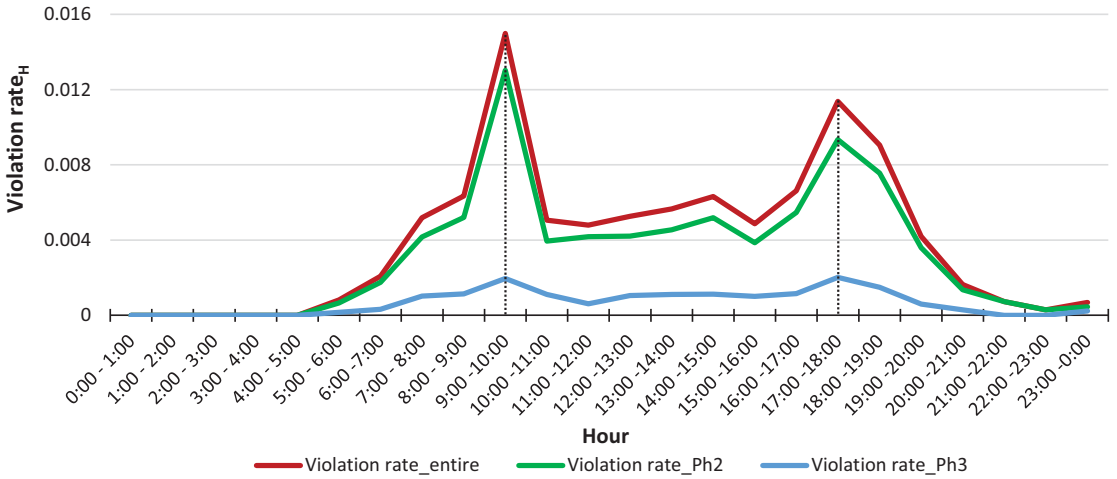
If we only consider the number of vehicle violations during Ph2 and Ph3, there is a possibility that, although the violation volume is high, the entire traffic volume is relatively high as well during the same period. Therefore, it is more appropriate to analyze the further violation rate as shown in Eq. (4.1). The analysis in this section also excludes the vehicles crossing the four LXs during the first 4 seconds.

$$\begin{aligned} Violation\ rate_H &= VV_H/VT_H \\ Violation\ rate_D &= VV_D/VT_D \end{aligned} \tag{4.1}$$

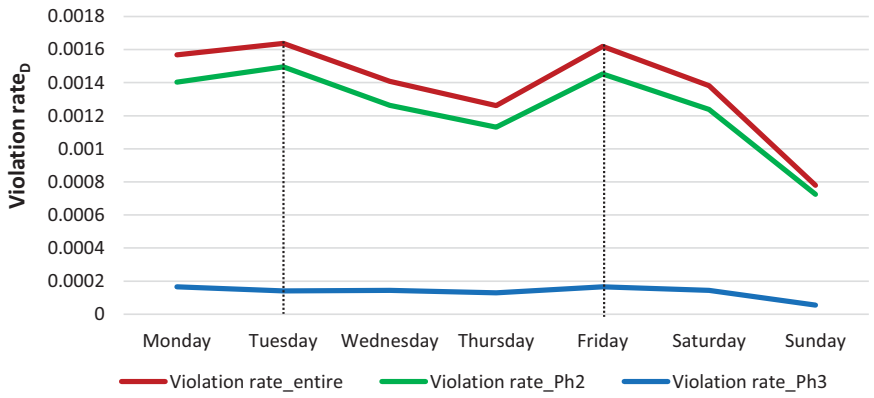
where $Violation\ rate_H$ represents the rate of violations during Ph2 and Ph3 in a given period of one hour; VV_H is the number of violating vehicles during one hour; VT_H is the total number of vehicles crossing the LX during the hour considered; $Violation\ rate_D$ represents the rate of violations during Ph2 and Ph3 in a given weekday; VV_D is the number of violating vehicles during a weekday; VT_D is the total number of vehicles crossing the LX during the day considered.

Fig. 4.6a illustrates the violation rate at the four LXs during Ph2 and Ph3 with regard to the hour in weekdays ($Violation\ rate_H$). Obviously, the violation rate in Ph2 is considerably higher than the violation rate in Ph3. Indeed, MV drivers know fairly well that the collision risk is higher as time advances during the LX closure cycle. One can also notice two peaks of the violation rate in weekdays, which fall into the period from 9:00 h to 10:00 h (0.015) in the morning and the period from 17:00 h to 18:00 h (0.011) in the afternoon, respectively. However, according to the statistics of the total number of crossing vehicles, the actual morning rush hour is from 8:00 h to 9:00 h, and the afternoon

rush hour is from 17:00 h to 18:00 h. It can also be noticed that the peak of the violation rate in the morning is higher than that in the afternoon. On the contrary, the actual traffic volume in the morning violation peak hour is lower than that in the afternoon violation peak hour.



(a) Violation rate according to the hour during Ph2 and Ph3 in weekdays



(b) Violation rate according to the weekday during Ph2 and Ph3

Fig. 4.6. Violation rate during Ph2 and Ph3 with regard to the weekday and the hour

Fig. 4.6b plots the violation rate during Ph2 and Ph3 according to the days of a week ($Violation\ rate_D$). The violation rate in Ph2 is still considerably higher than the violation rate in Ph3 during weekdays. A fact is found that there is one distinct peak of violation rate that appears on Friday (0.00162). After people get off work on Friday, they are likely to go out for their private activities because of the coming weekend. On the other hand, the violation rate on weekends declines dramatically. Particularly, a significant decrease in the violation rate shows on Sunday. Indeed, in general, people are less hurried during

the weekends than during workdays, but also because road traffic is more fluid during weekends at these LXs.

4.2.2.3 Analysis of the speed of violating vehicles

In this section, the speed of violating vehicles in Ph2 (excluding vehicle crossing during the first 4 seconds of Ph2) and Ph3 is examined. The speed of vehicles violating during Ph4 is quite low and dispersive, since vehicles would skirt the half barriers slowly to cross during this phase. Therefore, we do not analyze the speed of violating vehicles during Ph4.

We should recall that the road speed limit at these four LXs is 60 km/h, according to the field data from SNCF Réseau. Here, we analyze the violating vehicles after the first 4 seconds of Ph2 in the daytime (the number of violating vehicles recorded at night is very limited).

As shown in Fig. 4.7, the scatter plot shows the number of violating vehicles falling into the speed classifications concerned in terms of 2 and 5 time intervals in Ph2 and Ph3, respectively, (refer to section 4.2.2.1) during the daytime. The vehicle speed corresponds to the left vertical axis. Clearly, the number of violating vehicles with speed over 40 km/h decreases continuously to 0 in the last time slot of Ph3, as time advances. Moreover, the number of violating vehicles with speed between 10 km/h and 40 km/h decreases as time advances as well, but still keeps 10 validations in the last time slot of Ph3. Besides, violating vehicles with speed lower than 10 km/h is quite rare.

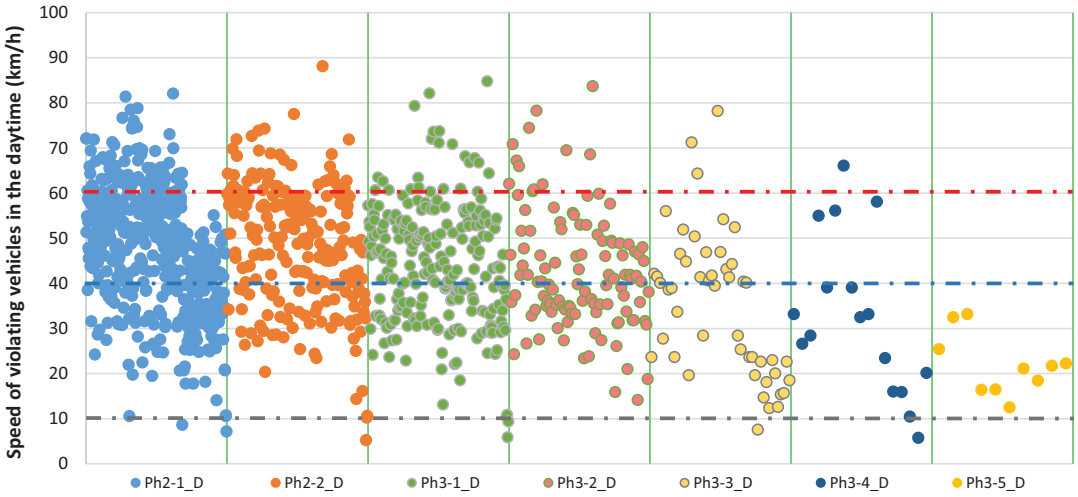


Fig. 4.7. Speed of violating vehicles and relative violation rate of vehicles with speed over 40 km/h in the daytime during Ph2 and Ph3

4.3 Motorist behavior analysis during Ph4

When analyzing LX accident scenarios, the scenario consisting in road vehicles bypassing the half barriers to cross the SAL2 LX when it is closed for the road traffic (zigzag) has been identified as a major scenario causing train-MV collisions [Ghazel 2017]. Therefore, the analysis of motorist behavior at SAL2 LXs during Ph4 of the LX closure cycle is important to scrutinize the major risky scenario causing LX train-MV accidents and allows for a detailed assessment of various aspects related to zigzag violations of motorists. This is our consideration that the study presented in this section is dedicated to discussing motorist behavior at SAL2 LXs during Ph4.

Due to the small sample size of zigzags in Ph4 observed at LX 55, 58 and 69[1], another seven SAL2 LXs, named according to their respective locations, i.e., LX 2 Toulouse, LX 4 Marnes-la-Coquette, LX 21 Pont-Sainte-Maxence, LX 71 Remaucourt, LX 82 Neufchâteau, LX 136 Choloy, LX 356 Caussade and one SAL4 LX (LX 425 Chaniers), were selected for our further field measurement campaign. Since our aim is to analyze motorist behavior at SAL2 LXs, the SAL4 LX is considered in our experiments for the purpose of comparing motorist responses to SAL2 and SAL4 LXs.

As shown in Fig. 4.8, 11 LXs are indicated on the French map. The red circular marks represent the accurate locations of the selected LXs and the railway stations in the vicinity of these LXs are also indicated by black triangle marks.

[1] Note that, the measurement data of LX 51 related to Ph4 are removed from our study due to the synchronization failure of Ph4 at LX 51.

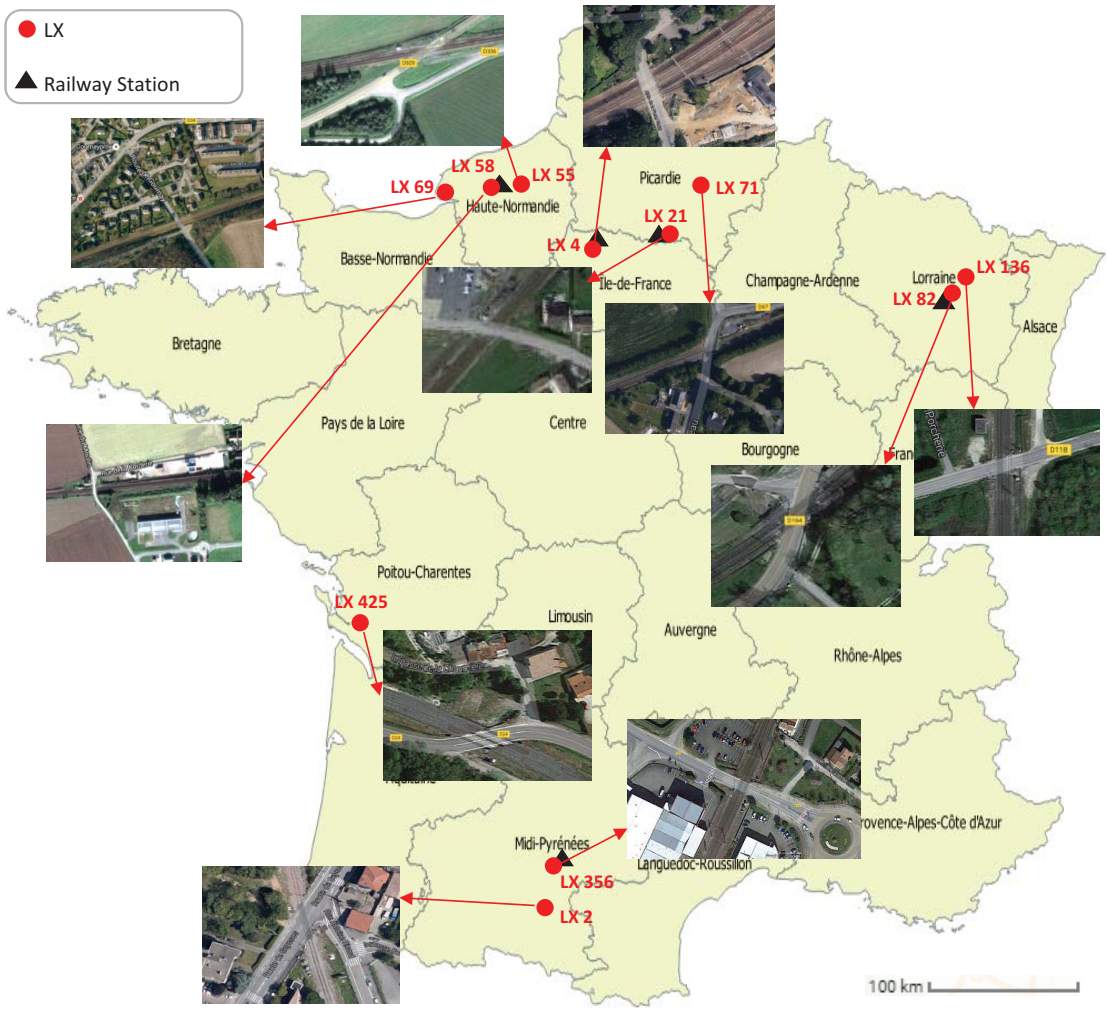


Fig. 4.8. Geographic information about 11 LXs

Table 4.2 shows various features (e.g., environment, infrastructure, equipment, railway line and road traffic involved, etc.) as well as their accident/incident statistics of the 11 LXs. It is worth noticing that when considering the number of accidents, LX 2 and LX 4 take the first place among the 11 LXs considered. While considering the sum of accidents and near misses, LX 55 takes the first place. In this stage of motorist behavior analysis related to Ph4, as for road traffic measurement, one can refer to section 4.2.1. In fact, during the period of observation, the total crossings recorded at the 10 SAL2 LXs were 461596, where there were 5678 crossings happened during the whole closure cycle and 116 zigzags happened during Ph4. Namely, the general average violation rate is 1.230% and the general average zigzag rate is 0.025%.

Table 4.2. Collisions/near misses and road/rail traffic volumes at the LXs from 1974 to 2014

LX	Type	# Accidents	# Near misses	# Average daily road traffic	# Average daily train traffic
LX 2	SAL2	5	7	4250	43
LX 4	SAL2	5	6	1184	124
LX 21	SAL2	0	8	570	77
LX 55	SAL2	3	114	21491	70
LX 58	SAL2	0	5	13938	89
LX 69	SAL2	0	1	10946	91
LX 71	SAL2	0	5	985	93
LX 82	SAL2	4	40	2500	94
LX 136	SAL2	3	8	3525	94
LX 356	SAL2	2	8	2892	59
LX 425	SAL4	1	4	5425	37

4.3.1 Behavioral analysis during Ph4

In this section, the violation rate related to prolonged Ph4 duration, zigzag occurrence related to LX location and dispersive Ph4 duration and finally, waiting queue and troop crossing phenomenon related to road traffic density and zigzag moment are discussed thoroughly.

4.3.2 The impact of prolonged Ph4 duration

The duration of Ph4 shows an important discrepancy from one closure cycle to another according to the actual speed of the train involved. In fact, the sensors responsible for announcing the train's approach to the local LX control system are implemented in such a way as to ensure that the LX closure is triggered to ensure a minimum given delay prior to the train arrival at the intersection zone. Therefore, the location of the announcement sensors which detect the train arrival is set according to the speed limit of the track section. Thus, according to the SNCF operation standard, the shortest time for train arriving at SAL2 intersection zone is 25 s. Nevertheless, since different train categories, such as freight and passenger trains (with different speeds) may pass on the section where the LX is located, such a control scheme leads to inevitable long closure durations in the case of slow trains. Train-MV collisions are most likely to occur when vehicles start to cross the LX five seconds before trains arriving since they would not have enough time to leave the intersection zone before the train arrives. The durations of Ph2 "Red Flash and

Siren" and Ph3 "Barriers Coming Down" at each LX are constant. Table 4.3 shows the duration of Ph2 and Ph3 at each LX. One can notice that the average duration of Ph2 and Ph3 altogether at these LXs is about 16 s. Namely, as shown in Fig. 4.9, collisions are most likely to occur after the first 4 s (16 s + 4 s = 20 s) of Ph4 (t > 4 s). Therefore, we focus more particularly on the motorist violations after the first 4 s of Ph4 in the following analysis.

Table 4.3. Duration of Ph2 and Ph3 at each LX

LX	Ph2 duration	Ph3 duration
LX 2	8	12
LX 4	7	8
LX 21	7	8
LX 55	6	9
LX 58	8	7
LX 69	7	9
LX 71	8	9
LX 82	7	8
LX 136	9	8
LX 356	8	8
LX 425	7	9

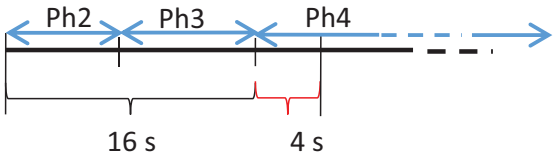


Fig. 4.9. The average violation rate of zigzags during Ph4 as Ph4 duration prolonged

In Ph4, violating vehicles skirt the half barriers to cross the LX, which represents a highly risky scenario. If one only considers the number of vehicle violations during Ph4, there is a possibility that, although the violation volume is high, the entire traffic volume is relatively high as well during the same period. Therefore, it is more appropriate to analyze the violation rate as expressed by Eq. (4.2).

$$Violation\ rate_{Ph4} = VV_M/(VD \times D_M) \tag{4.2}$$

where $Violation\ rate_{Ph4}$ represents the violation rate during Ph4, VV_M is the number of violating vehicles during a period D_M extending from the first 4 s of Ph4 up to the violation moment, VD is the road traffic density during Ph4. It should be noted that we

consider the value of the vehicle density during the one-hour time slot including the Ph4 considered to estimate the total number of vehicles waiting in front of the LX during D_M, since this number cannot be determined directly.

Table 4.4 shows the average road traffic density and the violation rate during the observation period at each LX, ranked according to the average road traffic density in descending order. Besides, the 11 LXs can be divided into 3 groups, i.e., G1, G2 and G3, according to the average road traffic density. The average road traffic density of the LXs in G1, G2 and G3 is more than 0.1, from 0.01 to 0.1 and less than 0.01, respectively. We aim to determine the impact of Ph4 duration on "zigzag" motorist behavior. Generally, one would conjecture that a higher rate of "zigzag" violations would appear as the duration of Ph4 is extended. A thorough analysis is carried out to validate whether this general speculation, commonly assumed, is correct.

Therefore, we adopt the average violation rate at those LXs having zigzags (cf. Table 4.4) to determine the overall trend of violation rate as time advances. The average violation rate is the average value of the samples of exact violation rate during each time interval involved, which can lessen the dispersion of exact violation rate values and make the trend clearer. In addition, as the number of zigzags recorded at night is very limited, we only consider those committed in the daytime (from 6:00 h to 21:00 h) exclusively. Hence, by reporting the average ratio of the number of recorded zigzags during D_M, to

Table 4.4. Information about average road traffic density and violation rate during the observation period at each LX

Group	LX	Average road traffic density (/s)	Violation rate (t > 4 s)	Near to railway station
G1	LX 55	0.247	0.134	No
G1	LX 58	0.167	1.957	Yes
G1	LX 69	0.126	1.060	No
G2	LX 356	0.062	0.414	Yes
G2	LX 2	0.054	0.995	No
G2	LX 425	0.053	–	No
G2	LX 82	0.037	–	Yes
G2	LX 136	0.018	–	No
G2	LX 71	0.017	–	No
G3	LX 4	0.009	0.259	Yes
G3	LX 21	0.005	0.715	Yes

"–" means that no zigzags are observed at these LXs during the experiment period.

the estimated number of vehicles during this period, and the duration of D_M, we determine the average violation rate distributed according to the time interval starting after the first 4 s of Ph4 up to the violation moment. As shown in Fig. 4.10, the global trend of the average violation rate is drawn as the duration of the waiting time interval extends. This curve attests that the violation rate during Ph4 decreases as Ph4 duration is prolonged (the longer the waiting time interval starting after the first 4 s of Ph4 up to the violation moment, the longer the Ph4 duration). This outcome demonstrates that the general intuitive speculation mentioned above that a higher rate of "zigzag" violations would appear as the duration of Ph4 is extended is not valid.

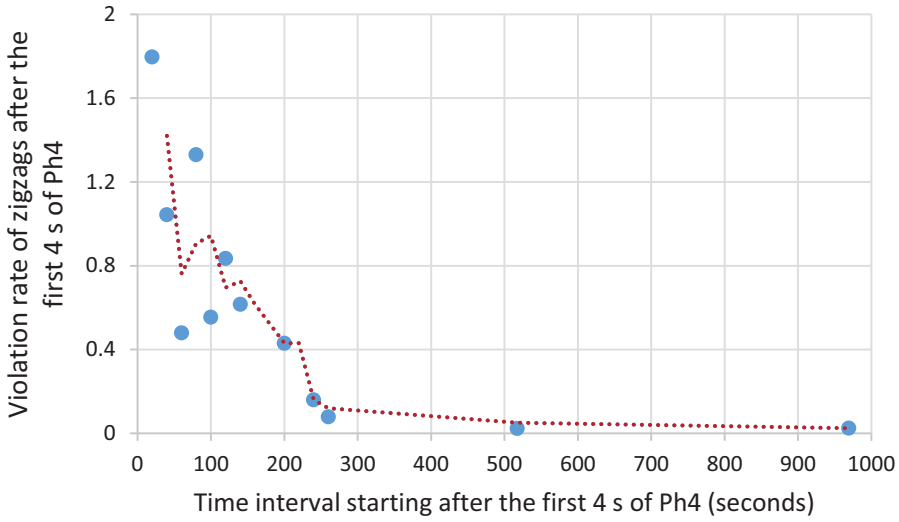


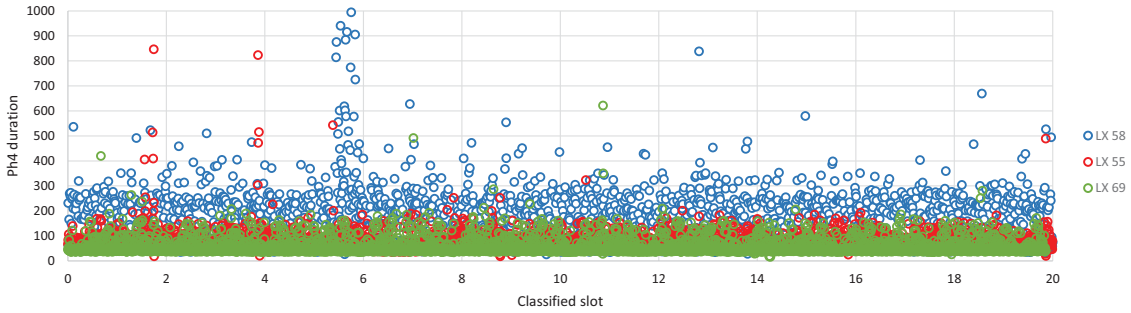
Fig. 4.10. The average violation rate of zigzags during Ph4 as Ph4 duration prolonged

4.3.3 The impact of LX location (near to railway station or not)

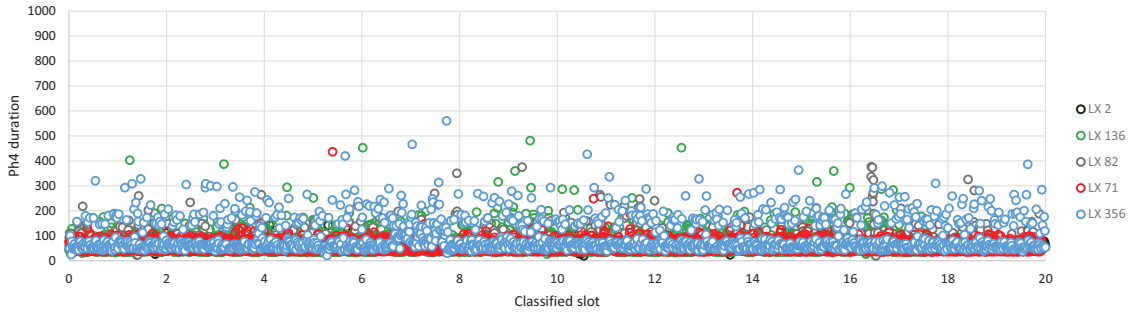
In order to further analyze the factors influencing zigzag occurrence, we scrutinize the Ph4 duration at each LX during the observation period. As shown in Fig. 4.11, Ph4 duration distribution during the observation period is drawn with regard to 20 equal classified slots.

As for the LXs in G1 (cf. Fig. 4.11a), the Ph4 duration distribution at LX 58 is quite dispersive. On the contrary, the Ph4 duration distributions at LX 55 and LX 69 are more centralized. As for the LXs in G2 and G3 (cf. Fig. 4.11b and Fig. 4.11c), the Ph4 duration distributions at LX 356 and LX 21 are relatively dispersive, compared with other LXs in these two groups, respectively. Referring to Table 4.4, these LXs with dispersive Ph4 durations are quite close to railway stations. Thus, some trains go through slowly or just stop above the LX announcement sensors for shortly stopping in railway stations nearby. Indeed, in some cases, the sensors are installed around or in the railway stations. This

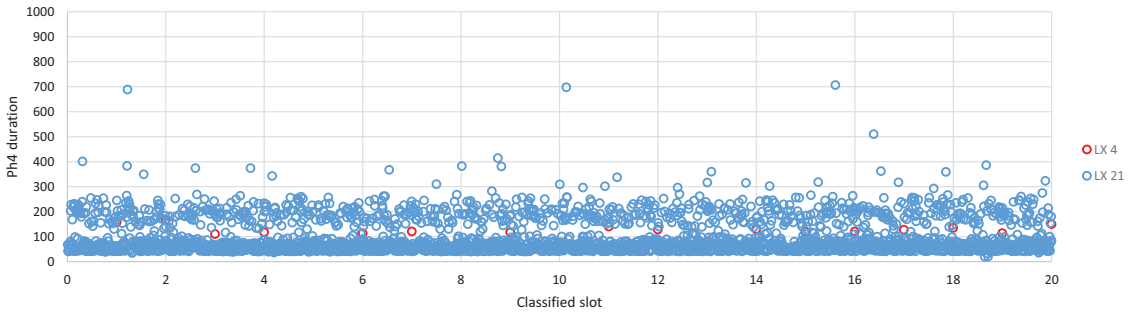
fact explains the dispersive Ph4 duration distribution.



(a) Ph4 duration distribution of LXs in G1



(b) Ph4 duration distribution of LXs in G2



(c) Ph4 duration distribution of LXs in G3

Fig. 4.11. Ph4 duration distribution with regard to classified slots

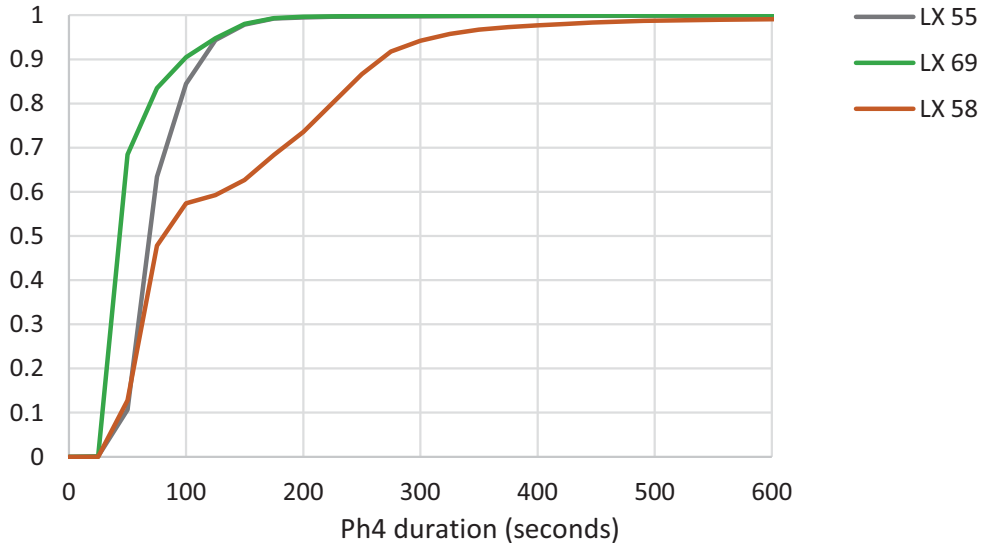
These dispersion characteristics can be indicated more clearly through Fig. 4.12 which shows the cumulative distribution of Ph4 duration at the considered LXs. As shown in Fig. 4.12a, it is worth noticing that, for the LXs in G1, the cumulative distribution curves of Ph4 duration at LX 55 and LX 69 climb dramatically during the first 100 s of Ph4, then reach 1 after 200 s. That means Ph4 durations at LX 55 and LX 69 are quite centralized within the first 100 s of Ph4. It is because, to a large extent, no railway stations are near to these LXs. However, the cumulative distribution curve of Ph4 duration at LX 58

shows inflection points at 100 s and 280 s, which indicates that Ph4 durations at LX 58 are relatively dispersive (from 100 s to 280 s). This is due to the fact that LX 58 is close to a railway station. More importantly, the violation rate at LX 58 is also the highest compared with that at LX 55 and LX 69.

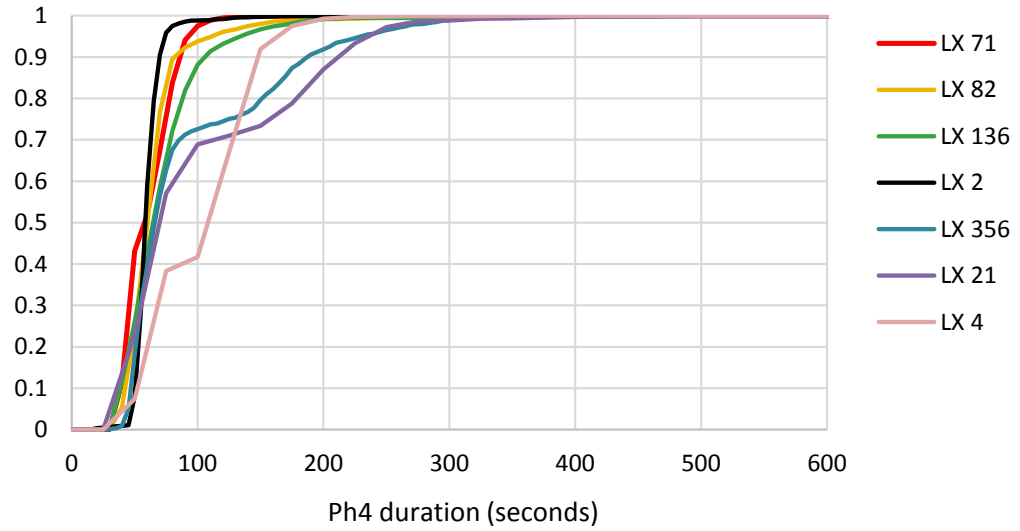
As for the LXs in G2 and G3 (cf. Fig. 4.12b), LX 21 has the most dispersive Ph4 duration followed by LX 356 and LX 4 in sequence, according to their cumulative distribution curves of Ph4 duration. One can still note that these 3 LXs are all close to railway stations. Moreover, LX 21 has the highest violation rate among these 3 LXs followed by LX 356 and LX 4 in sequence, accordingly. There is only one exception that LX 2 has a high violation rate but does not have dispersive Ph4 durations (while no railway station nearby). The surrounding of LX 2 will be further investigated in future works to explore other potential reasons for the high violation rate at this LX.

In order to thoroughly analyze the violation rate at each LX close to railway station, we extract the cumulative distribution (CD) curves of Ph4 duration at those LXs in Fig. 4.13. Moreover, the corresponding slope rates of CD between 0 s and 200 s of Ph4 are shown in table 4.5, which can reflect the dispersion of Ph4 duration at the five LXs. It is worth noticing that LX 58 has the most dispersive Ph4 duration (Slope rate = 0.0037) among these 5 LXs and LX 21 (Slope rate = 0.0044) takes the second place followed by LX 356 (Slope rate = 0.0046), LX 4 (Slope rate = 0.0050) and LX 82 (Slope rate = 0.0050) in sequence. It is worth noticing that LX 58 has the most dispersive Ph4 duration among these 5 LXs and LX 21 takes the second place followed by LX 356, LX 4 and LX 82 in sequence. On the other hand, as shown in table 4.4, the violation rates of the 4 LXs (except LX 82) follow the same order, that is LX 58 has the highest violation rate and then LX 21 takes the second place followed by LX 356 and LX 4 in sequence. There are no zigzags or dispersive Ph4 duration found at LX 82. After checking the railway infrastructure characteristics of LX 82, we have noted that the distance between LX 82 and the railway station is about 1.3 km, however, the distance between the LX announcement sensor and the LX is only around 700 m. Therefore, as for the LX announcement sensor on the odd track, it has a distance of about 2 km from the railway station; as for the LX announcement sensor on the even track, it still has a distance of about 600 m from the railway station. The railway speed limit of the track section including LX 82 is 150 km/h and the railway station also has a certain length (250 – 500 m). Namely, trains would not stop above the LX announcement sensors or slow down when they pass the LX announcement sensors on both opposite directions. That is why the railway station nearby does not cause dispersive Ph4 durations at LX 82. In summary, we can infer that motorists are more likely to commit zigzag violations at the LXs located close to railway stations, i.e., with dispersive Ph4 durations. Besides, the more dispersive the Ph4 durations, the

higher the violation rate of zigzags.



(a) The cumulative distribution of Ph4 duration at LXs in G1



(b) The cumulative distribution of Ph4 duration at LXs in G2 and G3

Fig. 4.12. The cumulative distribution of Ph4 duration

4.3.4 The impact of road traffic density

Waiting queues of vehicles in front of LXs can foster troop crossing phenomenon that represents a high-risk scenario. Logically, a waiting queue would be most likely to form

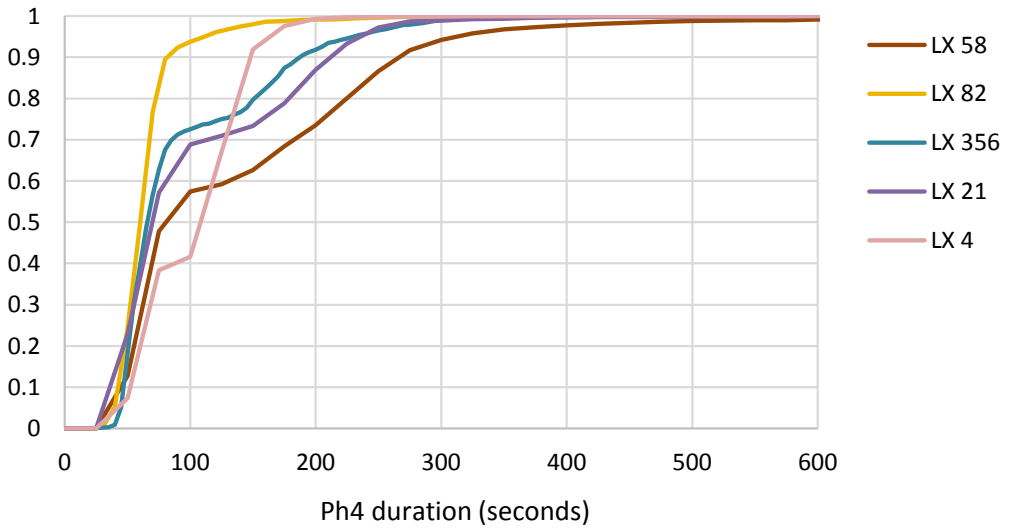


Fig. 4.13. The cumulative distribution of Ph4 duration at those LXs close to railway stations

Table 4.5. The slope rates of CD between 0 s and 200 s of Ph4 at those LXs close to railway stations

Group	LX	CD_200 s	Slope rate	Violation rate (t > 4 s)	Near to railway station
G1	LX 58	0.736	0.0037	1.957	Yes
G3	LX 21	0.870	0.0044	0.715	Yes
G2	LX 356	0.918	0.0046	0.414	Yes
G3	LX 4	0.990	0.0050	0.259	Yes
G2	LX 82	0.991	0.0050	–	Yes

"–" means that no zigzags are observed at the LX during the experiment period.

since the beginning of Ph3 "Barriers Coming Down"[2]. When vehicles arrive at an LX after a certain time since the beginning of Ph4, there are potentially some vehicles that have arrived earlier at the LX, which are waiting in front of the barriers. After a long time of waiting for the LX open cycle, some motorists may potentially lose their patience. Once the first vehicle in front of the LX commits zigzag crossing, the subsequent vehicles attempt to follow it closely. This scenario is called "troop crossing phenomenon".

We suppose that road traffic density has an impact on the occurrence of waiting queue and troop crossing. Therefore, we chose LX 55, LX 69, LX 2, LX 4 and LX 21 which showed relatively more zigzag occurrences compared with the other LXs.

The length of a waiting queue is reflected by the number of waiting vehicles accumu-

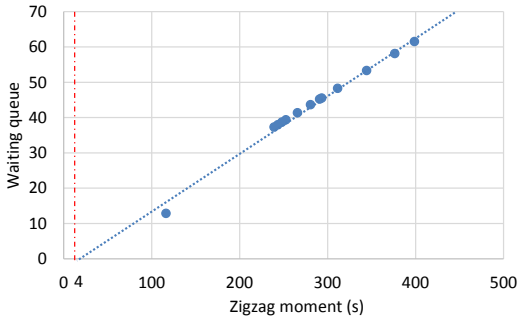
[2]according to statistics from SNCF Réseau

lated from the beginning of Ph3 up to the time moment of the zigzag occurring, which is expressed by Eq. (4.3).

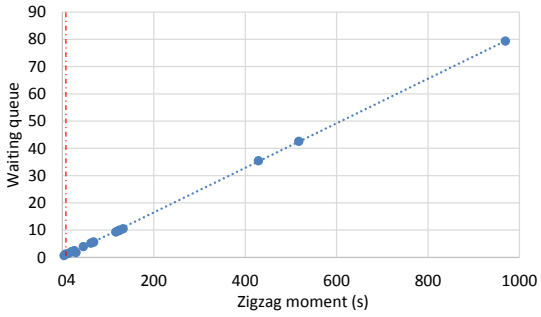
$$WQ = VD \times M \tag{4.3}$$

where WQ represents the length of waiting queue when the zigzag occurred; VD is the road traffic density during the closure cycle involved and M is the duration from the beginning of Ph3 up to the zigzag moment.

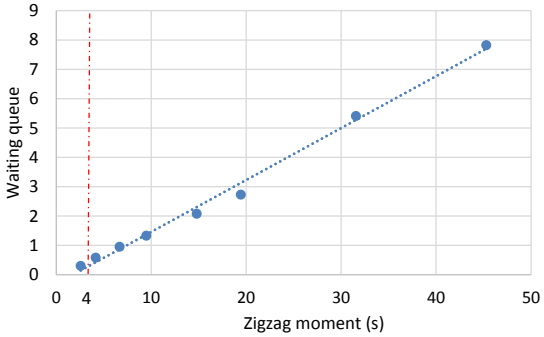
As shown in Fig. 4.14, waiting queues at these 5 LXs are drawn with regard to the time moment of zigzag occurring. The red dash line indicates the 4^{th} second of Ph4. The blue dotted line is an auxiliary indicator to show the dispersion of the scatter points in the chart. It is worth noticing that for LX 55 (cf. Fig. 4.14a) and LX 69 (cf. Fig. 4.14b) which show quite high road traffic density, their waiting queue length is almost concentrated on the blue line and then the cluster of waiting queue length points of LX 2 (cf. Fig. 4.14c) is next, which shows a little bit of dispersion around the blue line. However, for LX 4 (cf. Fig. 4.14d) and LX 21 (cf. Fig. 4.14e) which show significantly low road traffic density, their waiting queue length points are extremely dispersive.



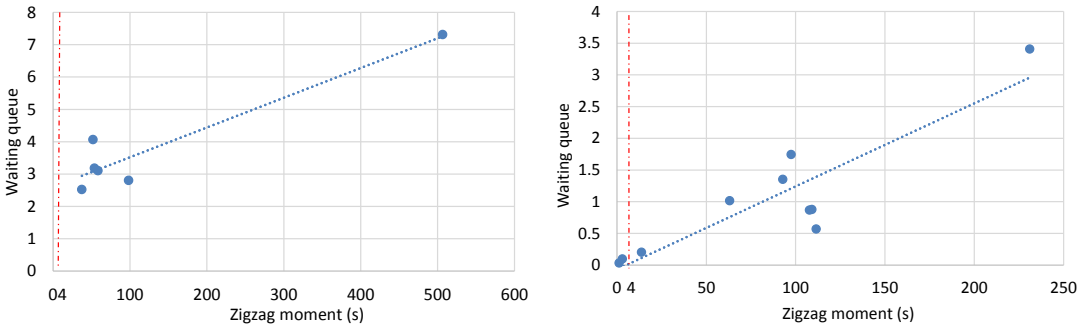
(a) The length of waiting queue at LX 55 in G1



(b) The length of waiting queue at LX 69 in G1



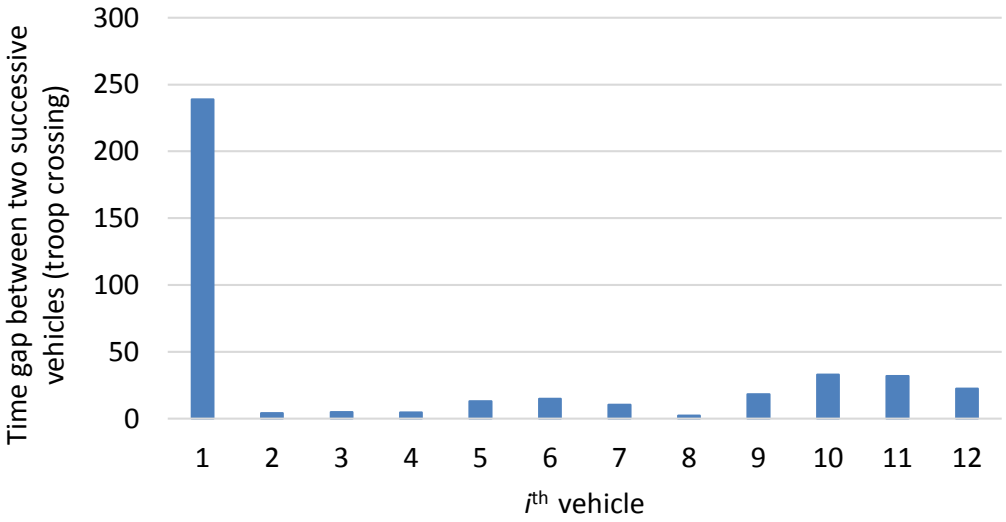
(c) The length of waiting queue at LX 2 in G2



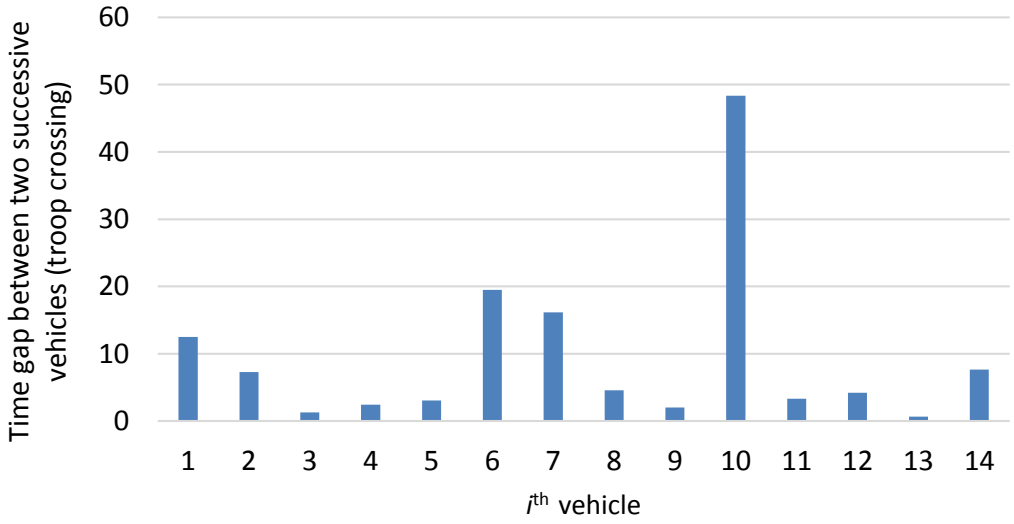
(d) The length of waiting queue at LX 4 in G3 (e) The length of waiting queue at LX 21 in G3

Fig. 4.14. The length of waiting queues at the 5 LXs

One can point out that the troop crossing phenomenon is more noticeable at LX 55 and LX 69 than at the other LXs, due to the highest road traffic density at LX 55 and LX 69 among these LXs. Fig. 4.15 shows the time gap between two successive zigzagging vehicles during a same Ph4, respectively at LX 55 and LX 69. This observation is clearer in Fig. 4.15a, indeed, the first vehicle commits zigzag crossing after a long waiting time of 240 s, and then the successive vehicles closely follow it to cross LX 55 as well. The average time gap between these following vehicles are 14 s and the shortest time gap is only about 4 s. A similar phenomenon emerges at LX 69. There are 3 times of troop crossing just after the 1^{st}, 6^{th} and 10^{th} vehicles commit zigzag crossing (cf. Fig. 4.15b). Moreover, the troop crossing phenomenon at LX 55 is more remarkable than that at LX 69 mainly because the road traffic density at LX 55 is higher than that at LX 69. Therefore, the aforementioned facts indicate that the higher the road traffic density at an LX, the more likely the waiting queues occur and the distribution of waiting queue length is more inclined to a linear distribution with regard to the duration from the beginning of Ph3 up to the zigzag moment. Furthermore, the troop crossing phenomenon is inclined to occur at LXs with significantly high road traffic density and the higher the road traffic density, the more distinct the troop crossing phenomenon.



(a) Troop crossing phenomenon at LX 55



(b) Troop crossing phenomenon at LX 69

Fig. 4.15. Troop crossing phenomenon at LX 55 and LX 69

4.4 Comparison of motorist responses to SAL2 and SAL4 LXs

LX 425, an SAL4 LX, is taken into account in our field observation in order to compare motorist responses to SAL2 and SAL4 LXs. For the purpose of making the results more

trustworthy, LX 356 and LX 2, which have the approximate road traffic density to that of LX 425 (differences less than 0.01), are selected for the comparison analysis in this section. The number of crossing vehicles during the closure cycle is analyzed to explore the trend of vehicle volume during various phases of the closure cycle. Since no zigzags are observed during Ph4 at LX 425 and due to the significant discrepancy of Ph4 durations from one closure cycle to another at each LX, only crossing vehicles during Ph2 and Ph3 are considered.

Turning to the data as shown in Table 4.3, the duration of Ph2 and Ph3 varies among the different LXs. The first 4 s of Ph2 is a special stage. The reason is explained as follows: when motorists observe the flashing lights and hear sirens, then, if it is possible to stop before the LX, they should brake. The reaction time, namely, the operating delay of the action guided by sense organs will take almost 1 s, according to ergonomics studies [Green 1967, Taylor 1967, Welford 1952]. In addition, the longest braking coordination time of motorized vehicles does not exceed 0.8 s. Then, the braking delay itself would be 2 s on average. Therefore, a vehicle crossing during the first 4 s of Ph2 will not be regarded as a violation in our study, only a vehicle crossing subsequently to the first 4 s of Ph2 should be considered as a violation, which may correspond to either an intentional violation or the inability of the motorist to stop his vehicle safely due to a high speed.

Therefore, as for the subsequent closure period after the first 4 s of Ph2, instead of considering the successive seconds in each phase (Ph2 and Ph3), we divide Ph2 (resp. Ph3) into 2 (resp. 5) equal time intervals at each LX and then, we can make the analysis in terms of phase halves: 2^{nd} and 3^{rd} (resp. phase quintiles: 1^{st}, 2^{nd}, 3^{rd}, 4^{th} and 5^{th}) in such a way as to be able to compare the road traffic volume in each time interval recorded at these LXs.

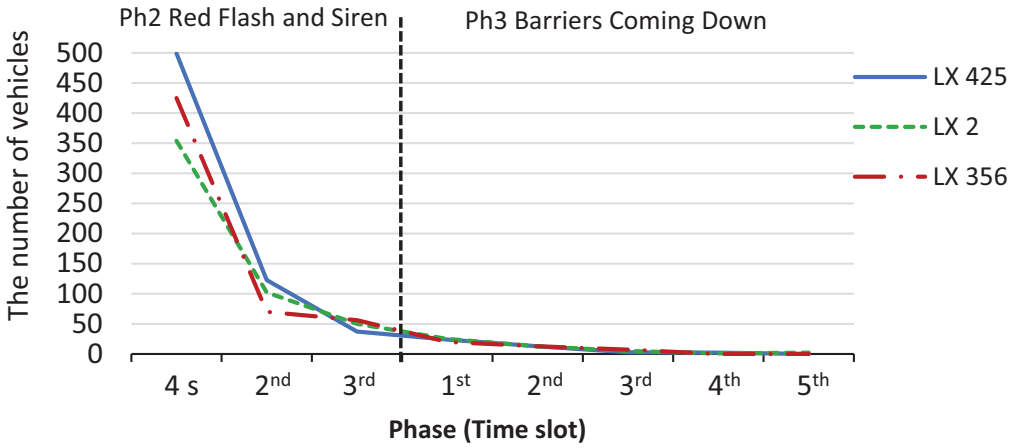
As shown in Fig. 4.16a, the number of vehicles crossing at the beginning of Ph2 at LX 425 is much bigger than that at LX 2 and LX 356. However, it decreases dramatically during the 2^{nd} stage of Ph2. The number of vehicles crossing in the 3^{rd} phase at LX 425 drops to the lowest among the 3 LXs. Moreover, one can notice that the scale of decrease at LX 425 is the biggest compared with the other SAL2 LXs.

Fig. 4.16b shows the normalized crossing ratio of vehicles during each phase of closure cycle at these 3 LXs. The normalized crossing ratio can eliminate the disparity of various road traffic densities at different LXs, which is expressed by Eq. (4.4).

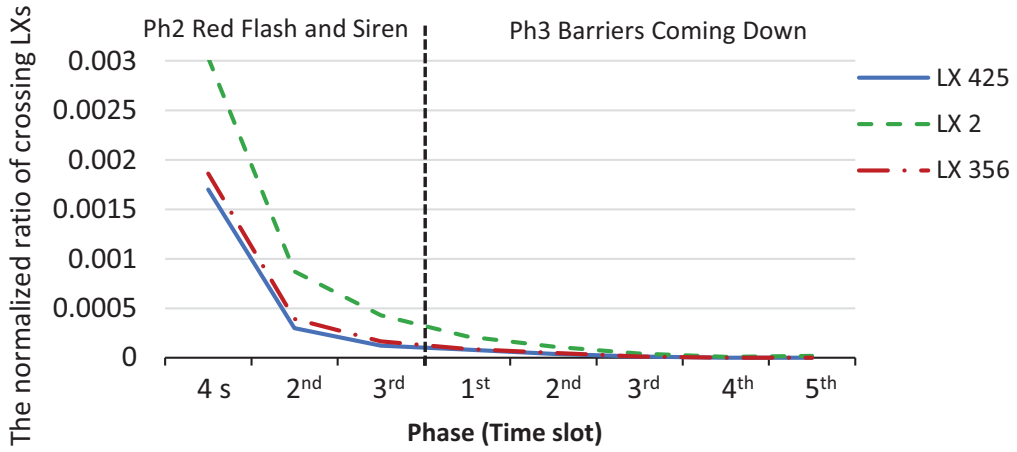
$$NC = Nb_V / Nb_{VT} \tag{4.4}$$

where NC represents the normalized crossing ratio during a time interval of the closure cycle at an LX; Nb_V is the number of crossing vehicles during the time interval considered, and Nb_{VT} is the total number of vehicles during the whole observation period at the LX

considered.



(a) The number of vehicles during each phase of closure cycle



(b) The normalized crossing ratio of vehicles during each phase of closure cycle

Fig. 4.16. Comparison of the number and normalized crossing ratio of vehicles at SAL2 and SAL4 LXs

According to Fig. 4.16b, obviously, the normalized crossing ratio during the closure cycle at LX 425 is the lowest among the 3 LXs. This indicates that motorists are more cautious when encountering closure cycles at an SAL4 LX and they scarcely cross an SAL4 LX during closure cycles. In fact, four half barriers of an SAL4 LX act as physical separators in four quadrants to effectively prevent zigzag crossing from both opposite directions. Another reason for motorists scarcely crossing an SAL4 LX is that, motorists know clearly that if they are blocked in the intersection zone, they will not be able to

escape from the closed SAL4 LX (unless they force the barriers at the exit site of the LX). Despite the new hazard of being blocked in the intersection zone (in the case of SAL4), the accident rate of SAL4 is 50% lower than SAL2 [SNCF 2017]. Further before/after measurements are needed to assess the efficiency of SAL4 LXs.

4.5 Discussion

According to the analysis in section 4.2.2, we can notice that, as time advances from Ph2 to Ph3, the global violation volume and violation rate of road vehicles both decrease dramatically. As for the impact of the schedule factor during Ph2 and Ph3, the peak of violation rate in the morning falls into the time slot between 9:00 h and 10:00 h. This is one hour later than the actual traffic rush hour in the morning. The peak of violation rate in the afternoon falls into the period from 17:00 h to 18:00 h. In addition, the peak of violation rate in the morning is higher than that in the afternoon. It is worth noticing that although the violation volume decreases as time advances along Ph2 then Ph3, it still shows a big number of violations at the beginning of Ph2. In order to reduce the violation volume from the beginning of Ph2, train-activated advance warning signs with flashers can be used for drawing motorists' attention in advance (a distance to the LX from which a safe stop can be achieved).

Based on the analysis of zigzag violations during Ph4 in section 4.3, we can summarize that the duration of Ph4, the existence of a nearby railway station and the road traffic density have significant impact on zigzag violations. It is recalled that during a prolonged Ph4, the longer the waiting time for vehicles, the lower the violation rate. This conclusion contradicts the general conjecture commonly assumed. Moreover, in terms of the global risk at a given LX, motorists are more likely to commit zigzag violations at the LX located close to a railway station. Indeed, the existence of a nearby railway station can foster the dispersion of Ph4 durations and, as illustrated earlier in the paper, the more dispersive the Ph4 durations at an LX, the higher the risk of zigzag violations at this LX.

Although the violation rate decreases as the time interval starting after the first 4 of Ph4 up to the violation moment prolongs, one cannot ignore the fact that the train-activated LX control system is susceptible to losing motorists' trust primarily due to variations in terms of Ph4 duration. Namely, it is possible for motorists to misunderstand the LX control system and commit zigzags when they have to wait for a long time without observing any train approaching, due to very slow train speed (e.g., they would think there is a failure in the LX control system, but it works well in fact). Reasonable and consistent warning time design is crucial to avoid zigzag violations during a long Ph4 [Ghazel 2017]. According to a report of [FHWA 1991], motorists would accept a shorter clearance time

at flashing lights and attempt to skirt barriers when the warning time exceeds 40-50 s. More importantly, second train coming active warning signs could help mitigate zigzag violations when Ph4 is prolonged by successive trains approaching. A short time gap between two successive approaching trains would cause a persistent closure of the LX after the first train has passed, since there is not enough time for LX control system to finish the activity of barriers uplifting before the second train arriving. Correspondingly, the Ph4 duration is prolonged by LX keeping closed. Such a kind of warning signs can avoid motorists losing patience when confronted with long LX closure cycles caused by successive approaching trains.

As for the impact of road traffic density, the higher the road traffic density at an LX, the more likely the waiting queue formation, hence, the troop crossing phenomenon is most likely to be boosted at LXs with significantly high road traffic density. In order to reduce the collision risk fostered by long waiting queues and troop crossing phenomenon, the installation of median separators and transforming SAL2 LXs into SAL4 LXs (Four-half barrier systems) [Ghazel 2014] is a further solution.

In fact, a median separator (used with an additional "U-Turn" or "Z-Turn" prohibition sign) in front of an SAL2 LX acts as a physical separator between opposing lanes of road traffic. Such a device effectively prevents zigzag violations with a visual cue intended to impede crossing to the opposing traffic lane. As shown in Table 4.6, the average length of waiting queue at the 4^{th} s of Ph4 at an SAL2 LX can be a preliminary guidance for designing the minimum length of a median separator at different LXs (the length of normal-size vehicle is considered: 3.8 – 4.3 m). However, before/after measurements are still needed to assess the efficiency of such solutions.

Table 4.6. The length of waiting queue (the number of waiting vehicles) at the 4^{th} s of Ph4 at each SAL2 LX

LX	Waiting queue length at the 4^{th} s of Ph4
LX 55	4.942
LX 58	3.334
LX 69	2.524
LX 356	1.242
LX 2	1.080
LX 82	0.732
LX 136	0.353
LX 71	0.347
LX 4	0.188
LX 21	0.105

According to the analysis in section 4.4, the SAL4 system is a significantly effective means to avoid zigzag violations of motorists. However, it is more costly than the other aforementioned technical solutions. Moreover, as mentioned in section 4.4, a new hazard will be introduced while using this device, that is vehicles could be trapped on the LX intersection zone in case of traffic jam when the four half (or two full) barriers come down. One solution would be to install additional obstacle detectors in order to reopen the exit barriers and order the train to brake. Such a protection system has been tested in the U.S. [Chadwick 2014]. Obstacle detectors should be intelligent enough to identify the difference between a car moving slowly and one that is stationary [Silmon 2010]. Any stationary vehicle staying for a long enough time on the LX intersection track can be assumed to be in trouble. In addition, according to the return on experience related to various LX safety strategies implemented in some countries [Anandarao 1998, ATC 2010, Davey 2005, Taylor 2008, Wullems 2011], a combination of education (aimed at changing road user behavior), enforcement and engineering measures is believed to lead to high safety level at LXs.

4.6 Summary

The study presented in this chapter offers a new perspective on motorist behavior during the closure cycles of LXs. The results obtained enable us to draw general conclusions and make recommendations for improving the whole LX safety. The analysis is carried out based on field measurement. Eleven SAL2 LXs and one SAL4 LX were selected for our field measurement. The data regarding road/railway traffic at these 12 LXs were collected, and then analyzed to derive the main features to characterize the behavior of motorists during LX closure cycles. The experimental equipment used in this study is at the same time reliable, practical, easy to install and economical. In summary, our ad-hoc installation is more convenient for our experiment, compared with cameras, automatic video surveillance or sophisticated video recording.

The main contributions of our study are that: the characteristics of motorist behavior are quantitatively and finely analyzed in various situations with regard to the various phases of closure cycle and some trends are established. Thus, corresponding targeted technical suggestions could be recommended accordingly, so as to improve LX safety. In detail, a comprehensive view is firstly shown through the analysis of the global trend of violating vehicle volume during Ph2 and Ph3 as time advances. Then the violation rate during Ph2 and Ph3 is analyzed in terms of hours and weekdays, respectively. Furthermore, a characterization of the violations according to vehicle speed has been carried out to analyze the decision-making process of vehicle drivers, as time passes in Ph2 and Ph3.

These thorough analyses could give us a deep understanding of vehicle drivers' behavior in different situations.

Finally, in order to explore the underlying reasons causing "zigzag" scenarios during Ph4, an in-depth study on the impact of influential factors, i.e., the duration of Ph4, the LX location (railway station nearby or not) and the road traffic density, on zigzag violations during Ph4 is carried out. Namely, the violation rate, the waiting queue and the troop crossing phenomenon during Ph4 are quantitatively and finely analyzed with regard to the above influential factors. Moreover, the distinction of motorist responses to SAL2 and SAL4 LXs is scrutinized to determine the efficiency of SAL4 LXs in terms of preventing zigzag violations. Thus, corresponding targeted technical suggestions can be recommended according to the specific analysis results pertaining to Ph4, so as to improve LX safety. Such thorough and novel analyses are rarely presented in existing works.

In summary, the aforementioned contributions of the study in this chapter are a direct response to the issues of understanding vehicle driver behavior with regard to different phases during the closure cycle of SAL2 LXs. The results obtained enable us to draw general conclusions and make recommendations for improving the whole LX safety.

Chapter 5

Advanced Statistical Accident Prediction Modeling

Sommaire

Overview

In this chapter, an advanced accident frequency prediction model, which enable to rank risky LXs accurately and identify the significant impacting parameters efficiently, is developed. Such a model takes into account the main parameters, namely, the average daily road traffic, the average daily railway traffic, the annual road accidents, the vertical road profile, the horizontal road alignment, the road width, the crossing length, the railway speed limit and the geographic region. The Nonlinear Least-Squares (NLS) method, Poisson regression method, negative binomial (NB) regression method, zero-inflated Poisson (ZIP) regression method and zero-inflated negative binomial (ZINB) regression method are employed to estimate the respective coefficients of parameters in the prediction model. Moreover, a validation process is performed based on various statistical and probabilistic means to examine how well the estimation of the model fits the reality. Besides, a comparison between the present model and two existing reference models is carried out to assess the efficiency of our model. The validation and comparison processes attest that the developed accident

prediction model with specified coefficients computed through the NLS method combined with NB distribution has statistic-based approbatory quality and relatively high predictive accuracy.

The work reported in this chapter has been published in "Safety Science" journal [Liang 2018d].

This chapter is structured as follows: section 5.1 exposes a general introduction to risk/accident prediction modeling techniques. The process of data collection and coding is given in section 5.2. In section 5.3, the general forms of the preliminary model and the advanced model developed are firstly presented; then, the coefficients associated with the parameters in the models are estimated using various regression approaches. In section 5.4, a thorough statistical analysis for examining the model quality and a comparison between the predictive accuracy of the specified regression models combined with the Poisson and NB distributions are then performed. In section 5.5, a comparison between our advanced model and some existing reference models is carried out to reflect the contribution of the current study. Subsequently, the impacts of various parameters on LX accident occurrence are discussed thoroughly in section 5.6. Finally, the main results and contributions of the present study are summarized in section 5.7.

5.1 Introduction

As investigated early in chapter 2, we recall here that L. E. Peabody and T. B. Dimmick developed Peabody-Dimmick Formula [Administration 2012] in 1941, which was used to estimate the number of accidents at railway-highway crossings within 5 years in 29 states in the U.S. and utilized through the 1950s. The average daily railway traffic T and the protection coefficient indicative of warning devices adopted P are considered in this formula. However, advances in both warning device technologies and LX design features quickly led to a no longer applicability of the predefined formula form and coefficients that reflected the conditions pertaining to LX accidents in 1941.

Subsequently, an evolutionary model of LX accident prediction called the New Hampshire Index [Oh 2006] was developed. The average daily road traffic, the average daily railway traffic and the protection factor indicative of the warning devices are taken into account.

The New Hampshire model is a relative formula which can be used to rank the importance of crossing upgrades. However, it is limited in that it does not predict the expected number of collisions, but only gives some indications about the priorities in terms of LX safety.

An accident prediction formula developed by the U.S. Department of Transportation (USDOT) in the early 1980s sought to overcome the limitations of earlier models [Chadwick 2014]. The initial collision prediction (prediction of collisions per year at a given LX), the exposure index (a variant of traffic moment) based on the product of highway and railway traffic, the index for the number of main tracks, the index for daily through trains during daylight, the index for highway paved, the index for maximum train speed,

the index for highway type and the index for highway lanes are considered in this model. The USDOT formula is the most commonly used model in the U.S. today. Although the formula is comprehensive, its current definition makes it difficult to identify or prioritize the design or improvement activities that will most effectively address LX safety-related issues, since it does not provide the magnitude of characteristics' contribution to the LX safety.

In Australia, a model called Australian Level Crossing Assessment Model (ALCAM) was developed [Committee 2012]. The ALCAM model takes into account how the physical properties at each LX site would affect human behavior, the LX control type, vehicle (or pedestrian) volumes, train volumes and the expected consequence of a collision which includes deaths and injuries involving both railway and roadway. The ALCAM has been applied across all Australian states and in New Zealand since 2003. However, it should be noticed that some LX physical properties considered in ALCAM show a high correlation between each other, which implies the existence of a kind of redundancy between the model inputs, and consequently a bias in terms of the outputs.

[Saccomanno 2003] validated several existing LX collision prediction models using the LX accident data of Canada. Various parameters, such as warning device, annual average daily traffic (AADT), surface width, train speed, number of tracks and daily number of trains, were considered in this study. The authors found that the expected collisions at LXs increased with traffic volume and that higher train speeds had a significant adverse impact on collisions at LXs with signs only, but not those with gates. [Hu 2010] used a generalized logit model to predict the level of accident severity at LXs in Taiwan and identify several significant impacting parameters. They reported that increasing number of daily trucks and number of daily trains were associated with a higher injury severity. [Khattak 2012] reported that the combined exposure of AADT and daily train traffic is more suitable than single AADT for LX collision prediction, because the probability of train-involved LX collisions in the absence of trains is zero. [Hao 2013] adopted Probit models to identify the contributing factors that influence the severity of injuries in accidents at U.S. LXs. This study showed that the peak hour, visibility, vehicle speed, annual average daily traffic, train speed, area type were significant.

In recent years, Poisson regression model, negative binomial (NB) regression model and other variants of the Poisson regression model [Lord 2010, Guikema 2012] have been preferred to deal with risk/accident statistics. [Oh 2006] adopted the expressions of the estimated expectation value $\hat{\lambda}$ as shown in Eq. (5.1) corresponding to the Poisson regression and NB regression models respectively when developing the U.S. LX accident prediction model. [Medina 2015] compared the USDOT model with the zero-inflated negative binomial (ZINB) model, in terms of accident prediction accuracy, using the LX accident data

from Illinois. The results of this study shows that the ZINB model has higher accuracy of prediction. [Lu 2016] employed the variants of Poisson regression model, for example, the zero-inflated Poisson (ZIP) model and the hurdle Poisson model, to deal with LX accident prediction involving the data in North Dakota. It should be noted that the expressions of estimated $\hat{\lambda}$ as shown in Eq. (5.1) are not appropriate in our current study, since they are limited to handling zero observations and some impacting variables should not be in the exponential form (e.g., there is a logical assumption that the predicted LX accident frequency should be 0 if the average daily railway traffic is 0). [Miranda-Moreno 2005] developed another model of $\hat{\lambda}$ as shown in Eq. (5.2). In this model, the product of the average daily road traffic V and the average daily railway traffic T (known as the conventional traffic moment, cf. section 5.3.1) is adopted.

$$\begin{aligned} \hat{\lambda}_{Poi} &= exp\left(\sum_{j=1}^{m} \beta_0 + \beta_j x_j\right) \\ \hat{\lambda}_{NB} &= exp\left(\sum_{j=1}^{m} \beta_0 + \beta_j x_j + \varepsilon\right) \end{aligned} \tag{5.1}$$

where β is the estimated regression coefficient; x is the impacting variable and ε is the gamma-distributed error in NB regression model.

$$\hat{\lambda} = (V \times T)^{\beta_1} exp\left(\sum_{j=1}^{m} \beta_j x_j + \sigma\right) \tag{5.2}$$

where $\sigma = \beta_0$ in Poisson regression model or $\sigma = \beta_0 + \varepsilon$ in NB regression model.

Based on these aforementioned investigations, it is clear that there is a strong need for an appropriate accident prediction model that should be comprehensive in terms of contributing factors to LX safety. Moreover, such a model should have high predictive accuracy. Therefore, in the present study, a new general accident prediction model is developed to predict the accident frequency at LXs.

5.2 Data sources and coding

The data to support our investigation come from a dedicated accident/incident database provided by SNCF Réseau (the French national railway infrastructure manager). SNCF Réseau has already investigated and recorded various attributes of LX accidents/incidents, railway and roadway traffic characteristics, surrounding characteristics of LXs. Therefore, the accident/incident data that cover SAL2 LXs in 21 geographical administrative regions in mainland France from 1990 to 2013 are obtained.

In the present study, an adequate sample is selected to include the data in the decade from 2004 to 2013, which provides reliable and sufficient information about both LX accidents and railway, roadway and LX characteristics. Namely, the selected LX inventory

presents the LX identification number, the LX location, the LX accident timestamp, the railway traffic volume, the road traffic volume, the LX dimension, the profile and alignment of the crossing road, etc. In total, there are 8332 public SAL2 LXs involved in our investigation. Using the LX identification number and the LX accident timestamp in the accident/incident database, the annual accident frequency at a given SAL2 is obtained. Then, a new database containing 10 years of data, which includes annual LX accident frequency, annual roadway accident statistics and railway, roadway and LX characteristics at a given SAL2, is created (again using the LX identification number as a common data index). The impacting parameters pertaining to LX accidents considered in our investigation should fulfill the following characteristics: (1) important in determining accident frequency, (2) more permanent in nature (e.g., sight obstruction noted as a problematic factor due to involved alterable construction topography, vegetation and other environmental elements) and (3) not accident-dependent [Austin 2002]. In fact, the combined database formed the basis of our investigation, and the statistical characterization of parameters considered in this investigation are shown in Table 5.1.

It should be noted that some minor data transformations in the combined database were necessary. Namely, the variables that have multiple non-numeric choices (e.g., profile, alignment) are transformed into singular indicator variables. On the other hand, numerical variables, such as the average daily road traffic, the average daily railway traffic, the railway speed limit, the LX width and the crossing length are used as they are without transformation. The region risk factor is determined by the general accident frequency per SAL2 in the region, while the road accident factor is determined by the ratio of the annual number of road accidents in a given year to the average number of road accidents per year over the period of 10 years considered. It is worth noticing that, by using the Spearman correlation checking [Borkowf 2002], we found that some other tested parameters were not significant (e.g., the road-rail track angle at a given LX) or highly correlated with some parameters considered in our analysis (e.g., the number of lanes at a given LX is highly correlated with the LX width). Overall, the data coding is shown in Table 5.2.

Table 5.1. Statistical characterization of parameters considered

Parameter	Description	Mean	Min	Max	StdDev
Railway traffic characteristics					
Average daily railway traffic	The average number of trains crossing the LX daily;	26.064	0.500	330	30.241
Railway speed limit	The maximum permission speed of train within the LX section;	92.460	5	160	42.383
Roadway traffic characteristics					
Average daily road traffic	The average number of road vehicles crossing the LX daily;	826.802	0.570	2.557e+04	1.781e+03
Annual road accidents	The number of road accidents in a given year;	7.050e+04	5.681e+04	8.452e+04	9.715e+03
LX characteristics					
Alignment	Horizontal road alignment shape: "straight", "curve" or "S";	N/A	N/A	N/A	N/A
Profile	Vertical road profile shape: "normal", "hump or cavity";	N/A	N/A	N/A	N/A
Length	The entering road width;	9.677	3	59	3.867
Width	The distance that road vehicles need to cross through the LX;	5.450	2	24	1.357
Region	The region of the LX considered;	N/A	N/A	N/A	N/A

Table 5.2. Parameters considered and data coding

Parameter	**Data coding**
Railway traffic characteristics	
Average daily railway traffic	Numerical, used directly;
Railway speed limit	Numerical, used directly;
Roadway traffic characteristics	
Average daily road traffic	Numerical, used directly;
Annual road accidents	Road accident factor: *National annual road accidents in a given year / National average road accidents per year over the period observed*;
LX characteristics	
Alignment	Alignment indicator: 0, 1 and 2 represent "straight", "curve" and "S", respectively;
Profile	Profile indicator: 0 and 1 represent "normal" and "hump or cavity", respectively;
LX width	Numerical, used directly;
Crossing length	Numerical, used directly;
Region	Region risk factor, highlighting the general LX-accident-prone region: *The number of SAL2 accidents over the observation period in the region considered / The number of SAL2 LXs in the region considered*;

For the detailed values of "Region risk factor" defined in Table 5.2, please refer to appendix A.2.

5.3 Accident prediction modeling

Based on some preliminary analyses, it is important to notice that five constraints need to be considered so as to develop the model for predicting annual accident frequency at a given SAL2:

- The predicted accident frequency should always be non-negative.
- It should be 0 if the average daily railway traffic is 0.
- It should be 0 if the average daily road traffic is 0.
- It should be 0 if the annual road traffic accidents are 0.
- The model should be time-dependent, i.e., it should reflect the variation of accident frequency as time advances.

Therefore, some differences can be noticed in our accident prediction model compared with the models shown in Eq. (5.1).

5.3.1 Preliminary accident prediction model

For the preliminary accident prediction model, we considered only three parameters in Table 5.2, which are the average daily railway traffic, the average daily road traffic and the annual road accidents. The preliminary model is developed as follows:

$$\lambda_{10P} = K \times F_{RAcc} \times V^a \times T^b \tag{5.3}$$

where λ_{10P} represents the annual accident frequency at a given SAL2 during the period of 10 years considered; K is the constant coefficient; F_{RAcc} is the road accident factor at national level; V is the average daily road traffic and T is the average daily railway traffic. Here, F_{RAcc} is a time-dependent variable which can reflect the variation of annual road accidents as time advances.

We should recall here that the conventional formula of the traffic moment is given as: Traffic moment = Road traffic frequency × Railway traffic frequency [Liang 2017c]. However, based on some previous analyses [Liang 2018d], we adopt a variant called "corrected moment", or CM for short. $CM = V^a \times T^b$, where $a + b = 1$ and the best value of a in terms of fitting is computed to be $a = 0.354$ according to the previous statistical analysis

performed by SNCF Réseau [SNCF Réseau 2010], since railway traffic has a more marked impact on LX accidents than road traffic. Therefore, we consider $(V^{0.354} \times T^{0.646})$ as an integrated parameter that reflects the combined exposure frequency of both railway and road traffic. One can notice that Eq. (5.3) can be rewritten as $\lambda_{10P} = K \times RM$, where $RM = F_{RAcc} \times V^{0.354} \times T^{0.646}$. Thus, this model can be regarded as a linear model with respect to the composite parameter RM. The Ordinary Least-Squares (OLS) method is employed to estimate coefficient K. As shown in Fig. 5.1, K is estimated as 1.319e-04 ($t - statistic = 33.72 > 1.96$ corresponding to a 95% confidence level).

Fig. 5.1 indicates that this preliminary model shows that, for high values of corrected moment, there is a significant deviation between observed accident frequencies and predicted accident frequencies at SAL2 LXs. Therefore, further statistical analysis is carried out to evaluate the quality of the transformed linear model. In this case, we make group classification, which means that the data set is divided into 100 groups with the same number of samples in each group. Then, the mean value of λ_{10P} and RM of each group are computed respectively to generate a linear relationship between the group-mean λ_{10P} and the group-mean RM. Hence, we can adopt "Residuals vs. Fitted" graph, "Normal Q-Q" graph, "Scale-Location" graph and "Residual vs. Leverage" graph [Anscombe 1973, Chatterjee 2015] to check the linearity, the normal distribution of residuals, the homoscedasticity [Jarque 1980] and the abnormal values of the model, respectively. These four graphs pertaining to our group classification analysis are shown in Fig. 5.2. For an idealized linear model: 1) the red line in Fig. 5.2a should be a horizontal line at 0 and residuals should randomly distribute around this line; 2) the standardized residuals shown in Fig. 5.2b should fall in the 45° direct line, which can attest the normal distribution of residuals; 3) the red line shown in Fig. 5.2c should be a horizontal line at a certain value and square roots of standardized residuals should randomly distribute around this line; 4) Fig. 5.2d can identify abnormal values and significant values which have an important impact on the model fitting, through Cook's distance. One can notice that residuals in the groups with big ID (i.e., 95, 99 and 100) are significant. These residuals correspond to SAL2 LXs with high corrected moment in our data set. Therefore, the statistical test results attest the significant deviation between the observed accident frequencies and the predicted accident frequencies at SAL2 LXs having high corrected moment. Through checking the accident/incident data, these SAL2 LXs with high corrected moment are correspondingly accident-prone LXs in general. We conjecture that this preliminary accident prediction model is not appropriate to predict the annual accident frequency at SAL2 LXs with high corrected moment. A thorough analysis needs to be performed to develop an advanced model which can predict the annual accident frequency at a given SAL2 more accurately. Meanwhile, the model should consider more impacting variables.

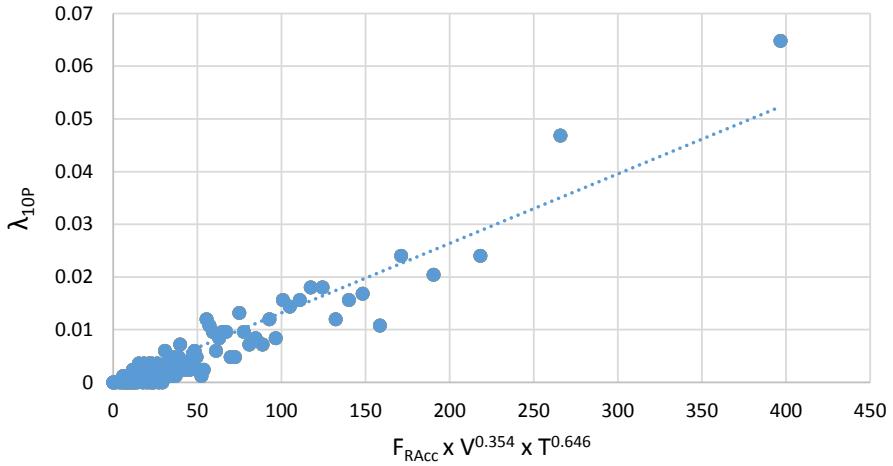


Fig. 5.1. The preliminary accident frequency prediction model λ_{10P} vs. $(F_{RAcc} \times V^{0.354} \times T^{0.646})$

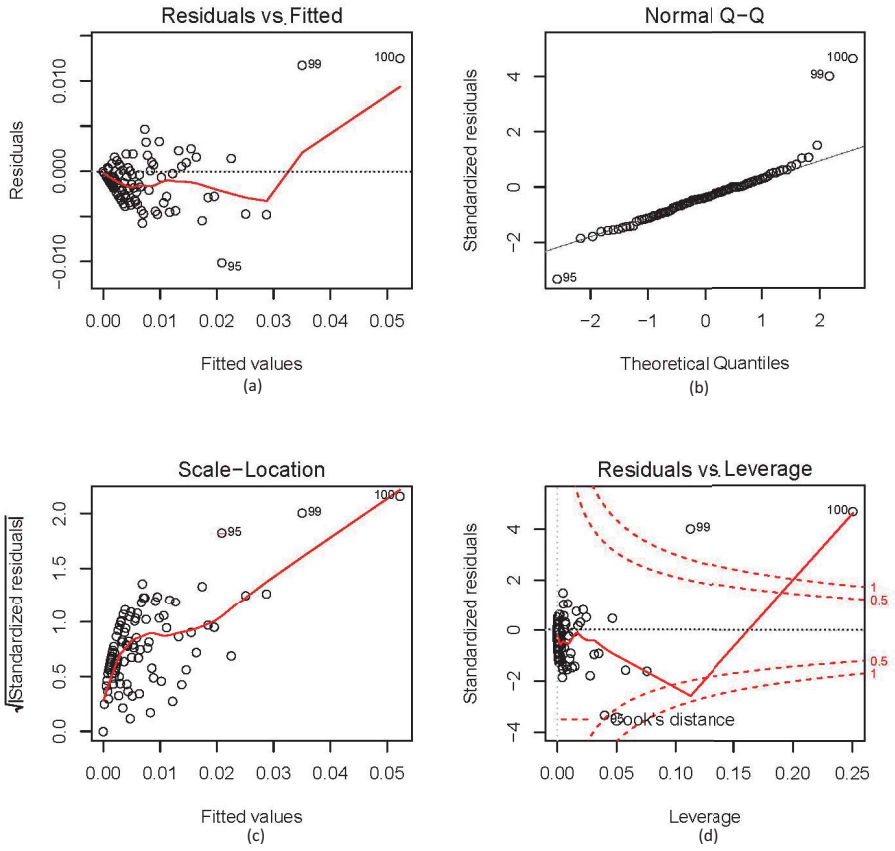


Fig. 5.2. Statistical evaluation of the model quality: (a) Residuals vs. Fitted, (b) Normal Q-Q, (c) Scale-Location, (d) Residual vs. Leverage

5.3.2 Advanced accident prediction model

The developed advanced model takes into account various variables as interpreted in Table 5.2. The general form of the model is shown as follows:

$$\begin{aligned}\lambda_{10Y} =& K \times F_{RAcc} \times (V^a \times T^b) \times exp(C_{Profile} \times I_{Profile} + C_{Align} \times I_{Align} + C_{Wid} \times Wid \\ &+ C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg})\end{aligned} \tag{5.4}$$

where λ_{10Y} represents the annual accident frequency at a given SAL2 for a period of 10 years; K is the constant coefficient; F_{RAcc} is the road accident factor which reflects the variation of annual road accidents as time advances (a time-dependent variable); V is the average daily road traffic; T is the average daily railway traffic; $I_{Profile}$ and $C_{Profile}$ are respectively the profile indicator and its corresponding coefficient; I_{Align} and C_{Align} are respectively the alignment indicator and its corresponding coefficient; Wid and C_{Wid} are respectively the LX width and its corresponding coefficient; $Leng$ and C_{Leng} are respectively the crossing length and its corresponding coefficient; RSL and C_{RSL} are respectively the railway speed limit and its corresponding coefficient; F_{Reg} and C_{Reg} are respectively the region factor and its corresponding coefficient (cf. Table 5.2).

Note that appropriate higher orders and interaction terms of covariates can be included in the model expressed by Eq. (5.4) without difficulty, due to the use of an exponential form [Miaou 1994]. Moreover, this model does not only rank risky LXs accurately, but also allow for identifying significant parameters efficiently.

5.3.2.1 Regression approaches

In this section, several regression approaches are adopted to estimate the coefficients associated with the parameters of our model. The Nonlinear Least-Squares (NLS) technique and Gauss-Newton algorithm [Madsen 2004] are firstly considered to estimate the variable coefficients in our model. Considering a fitting model function $y = f(x, \boldsymbol{\beta})$, where variable x depends on a vector of l parameters: $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \ldots,\ \beta_l)$. The goal is to find the vector $\boldsymbol{\beta}$ which can let the model function fit best the actual observed data in the least-squares sense. In other words, minimize the sum of residual squares S expressed as follows:

$$S = \sum_{i=1}^{m} r_i^2,\ m \geq l \tag{5.5}$$

where r_i is the residual between the fitting model estimation and the actual observation, $r_i = y_i - f(x_i, \boldsymbol{\beta})$.

The minimum value of S is obtained by solving the gradient function $\partial S/\partial\beta_j = 0$, i.e.:

$$\begin{aligned} &\partial S/\partial\beta_j = 2\textstyle\sum_i r_i \partial r_i/\partial\beta_j = 0 \\ &\beta_j \approx \beta_j^{k+1} = \beta_j^k + \Delta\beta_j \end{aligned} \tag{5.6}$$

where k is the iteration number and $\Delta\beta_j$ is the shift parameter.
At each iteration step, the model is linearized by approximation to a first-order Taylor series expansion about $\boldsymbol{\beta}^k$:

$$f(x_i,\boldsymbol{\beta}) \approx f(x_i,\boldsymbol{\beta}^k) + \sum_{j=1}^{l}(\beta_j - \beta_j^k)\partial f(x_i,\boldsymbol{\beta}^k)/\partial\beta_j \approx f(x_i,\boldsymbol{\beta}^k) + \sum_{j=1}^{l} J_{ij}\Delta\beta_j \tag{5.7}$$

where J_{ij} is the element of Jacobian matrix $\boldsymbol{J}$ and $\partial r_i/\partial\beta_j = -J_{ij}$.
Therefore, r_i can be rewritten as:

$$\begin{aligned} &r_i = \Delta y_i - \textstyle\sum_{s=1}^{l} J_{is}\Delta\beta_s \\ &\Delta y_i = y_i - f(x_i,\boldsymbol{\beta}^k) \end{aligned} \tag{5.8}$$

By substituting the above expressions into the gradient equation in Eq. (5.6), we obtain the normal equation and its matrix notation:

$$\begin{aligned} &\textstyle\sum_{i=1}^{m}\sum_{s=1}^{l} J_{ij}J_{is}\Delta\beta_s = \sum_{i=1}^{m} J_{ij}\Delta y_i \\ &(\boldsymbol{J}^T\boldsymbol{J})\Delta\boldsymbol{\beta} = \boldsymbol{J}^T\Delta\boldsymbol{y} \end{aligned} \tag{5.9}$$

For an NLS model, S should be modified as follows:

$$S = \sum_{i=1}^{m} W_{ii} r_i^2,\ m \geq l \tag{5.10}$$

Therefore, the matrix notation of normal equation for an NLS model is expressed as follows:

$$(\boldsymbol{J}^T\boldsymbol{W}\boldsymbol{J})\Delta\boldsymbol{\beta} = \boldsymbol{J}^T\boldsymbol{W}\Delta\boldsymbol{y} \tag{5.11}$$

These aforementioned equations form the basis of the Gauss-Newton algorithm for solving an NLS problem.

In fact, the Poisson regression model shown as Eq. (5.12) is a natural choice for modeling accident occurrence.

$$Poi(X=k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0,1,2,\ldots \tag{5.12}$$

where $Poi(X=k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$ and λ is the expectation value of the number of accidents.

However, [Chang 2005] indicates that accident frequency is likely to be over-dispersed (cf. Eq. (5.13)) and suggests using the negative binomial (NB) regression model as an alternative to the Poisson model.

$$VAR(X) \begin{cases} = E(X) \\ > E(X), \text{ over-dispersed} \\ < E(X), \text{ under-dispersed} \end{cases} \tag{5.13}$$

The NB model as a special case of Poisson-Gamma mixture model is a variant of the Poisson model designed to deal with over-dispersed data [Buddhavarapu 2016, Lord 2010, Utkin 2015]. The over-dispersion could come from several possible sources, e.g., omitted variables, uncertainty in exposure data, covariates or non-homogeneous LX environment [Miaou 1994]. The NB model considered in this study has the following expression:

$$P_{NB}(X = k) = \frac{\Gamma\left(k + \frac{1}{\alpha}\right)}{\Gamma\left(k+1\right)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k}, \; k = 0, 1, 2, \ldots \tag{5.14}$$

where $P_{NB}(X = k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$; λ is the expectation value of the number of accidents and α is the dispersion parameter.
The relationship between the mean value and the variance in the NB model is given as follows:

$$VAR(X) = E(X) + \alpha E(X)^2 \tag{5.15}$$

If $\alpha > 0$, there is an over-dispersion; if $\alpha < 0$, there is an under-dispersion and in the case where $\alpha = 0$, the NB model reduces to the Poisson model.

In practice, the count data may contain extra zeros relative to the Poisson or NB distribution. In this case, the ZIP or ZINB regression model is useful for analyzing such data [Ridout 2001]. The ZIP model is expressed as follows:

$$P_{ZIP}(X = k) = \begin{cases} \omega + (1-\omega)exp(-\lambda), \; k = 0; \\ (1-\omega)exp(-\lambda)\lambda^k/k!, \; k > 0 \end{cases} \tag{5.16}$$

where $P_{ZIP}(X = k)$ is the probability of k accidents occurring, $k \in \mathbb{N}$; λ is the expectation value of the number of accidents and $log\left(\frac{\omega}{1-\omega}\right) = z'\gamma$ is the ZI link function that z' is the ZI covariate and γ is the corresponding ZI coefficient. The mean value and variance of ZIP model are $E(X) = (1-\omega)\lambda$ and $VAR(X) = (1-\omega)\lambda(1+\omega\lambda)$.

The ZINB model is expressed as follows:

$$P_{ZINB}(X=k)=\begin{cases}\omega+(1-\omega)(1+\alpha\lambda)^{-1/\alpha}, & k=0;\\ (1-\omega)\dfrac{\Gamma\left(k+\dfrac{1}{\alpha}\right)}{\Gamma(k+1)\Gamma\left(\dfrac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha}\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k}, & k>0\end{cases} \tag{5.17}$$

where $P_{ZINB}(X=k)$ is the probability of k accidents occurring, $k\in\mathbb{N}$; λ is the expectation value of the number of accidents. The mean value and variance of ZINB model are $E(X)=(1-\omega)\lambda$ and $VAR(X)=(1-\omega)\lambda(1+\omega\lambda+\alpha\lambda)$. The ZINB reduces to the ZIP in the limit $\alpha\to 0$.

However, the NB and ZINB models are limited to handling under-dispersed data ($\alpha<0$) [Lord 2010]. If the dispersion parameter α is set as a negative value to try to handle under-dispersion issue, they would no longer be an NB or ZINB model and lead to unreliable estimation, especially when the sample mean is low and the sample size is small [Lu 2016]. That is why [Oh 2006] proposed the Gamma model to handle under-dispersed samples. The Gamma model is given as follows:

$$P_G(X=k)=Gamma(\beta k,\lambda)-Gamma(\beta(k+1),\lambda) \tag{5.18}$$

where $P_G(X=k)$ is the probability of k accidents occurring, $k\in\mathbb{N}$; λ is the expectation value of the number of accidents and β is the dispersion parameter. If $\beta<1$, there is an over-dispersion; while if $\beta>1$, there is an under-dispersion and if $\beta=1$, the Gamma model reduces to the Poisson model. However, the Gamma model shown in Eq. (5.19) is limited to the time-dependent observation assumption and zero observations, since general $\Gamma(x)$ restricts discrete responses to positive values.

$$Gamma(\beta k,\lambda)=\begin{cases}1, \text{ if } k=0\\ \dfrac{1}{\Gamma(\beta k)\int_0^{\lambda}u^{\beta k-1}e^{-u}du}, \text{ if } k>0\end{cases} \tag{5.19}$$

According to the above discussion, the restriction between variance and mean value is significant to identify an appropriate regression model. Therefore, we firstly adopted group classification to make preliminary variance analysis, which is that the annual accidents at a given SAL2 during the 10 years were divided into 100 groups with the same number of samples in each group. Then, the mean value and variance of accidents in each group were computed respectively to analyze the relationship between the group variance and the group mean value. The variance analysis is shown in Fig. 5.3. It seems that there is a slight over-dispersion of the data set since the variance $VAR(n)$ is a bit bigger than the mean $E(n)$ ($VAR(n)\approx 1.0171E(n)$). Hence, since the mean value and the variance are

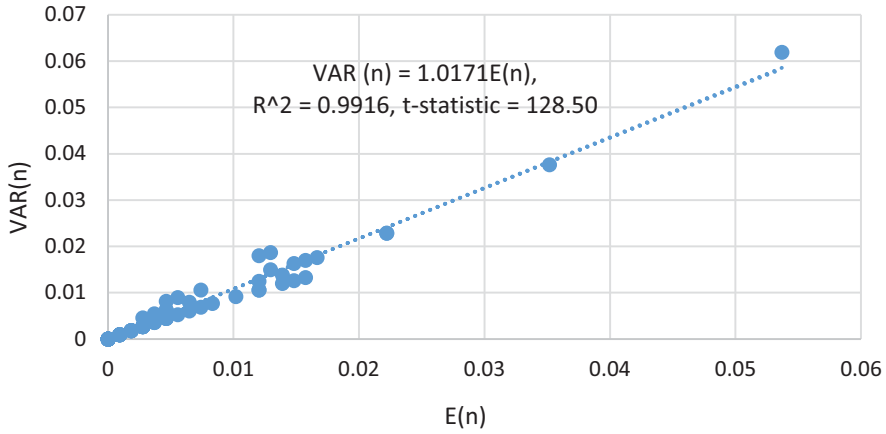


Fig. 5.3. Constraint between the group variance and the group mean value of annual accidents at SAL2 LXs

very close to each other, we performed meticulous analyses to assess the NLS regression, the Poisson regression, the ZIP regression, the NB regression and the ZINB regression methods with regard to SAL2 LXs in our accident database so as to identify which model is more effective.

5.3.2.2 Regression modeling results

NLS regression

When applying the NLS regression, the form of λ_{10Y} is given by Eq. (5.4). The estimated coefficients computed by NLS regression are provided in Table 5.3. A $|t - statistic| > 1.96$ is introduced to identify the significant parameters corresponding to a 95% confidence level. As a result, the average daily railway traffic, the railway speed limit, the average daily road traffic, the annual road accidents, the road alignment, the LX width, the crossing length and the LX-accident-prone region have been shown to have significant and positive influence on SAL2 accident frequency. However, the test shows that the road profile is not a significant factor ($|t - statistic| = 0.635 << 1.96$), thus the impact of road profile could be neglected. Moreover, the coefficients of the considered variables with the exponential form can reflect the sensitive degrees of the SAL2 accident frequency to these variables respectively. According to these sensitive degrees (rank indicated in brackets), the LX-accident-prone region factor is the most sensitive contributor among these variables.

Table 5.3. Results of the λ_{10Y} NLS regression model

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K	2.703e-05	5.078e-06	5.322	×
$I_{Profile}$	$C_{Profile}$	3.626e-02	5.706e-02	0.635	
I_{Align}	C_{Align}	3.427e-01 (2)	2.942e-02	11.648	×
Wid	C_{Wid}	9.847e-02 (3)	1.494e-02	6.589	×
$Leng$	C_{Leng}	2.084e-02 (4)	4.284e-03	4.865	×
RSL	C_{RSL}	3.089e-03 (5)	7.586e-04	4.072	×
F_{Reg}	C_{Reg}	4.962e-01 (1)	1.722e-01	2.882	×

In order to assess the predictive accuracy of accident occurrence estimated by the NLS regression model λ_{10Y} combined with the NB and ZINB distributions (cf. section 5.4.3), we adopt the Maximum Likelihood Estimation (MLE) method to estimate the dispersion parameter α of the data set [Dai 2013]. As expressed by Eq. (5.20) and Eq. (5.21), the values of α in NB and ZINB distributions are estimated respectively, using R language to solve $\partial l/\partial\alpha = 0$.

$$l(\alpha)_{NB} = ln(\prod_i^n P_{NB}(X_i = y_i)) = \sum(y_i ln(\lambda_i) - (y_i + \alpha^{-1})ln(1+\alpha\lambda_i) + \sum_{v=0}^{y_i-1} ln(1+\alpha v)) \tag{5.20}$$

$$l(\alpha)_{ZINB} = ln(\prod_i^n P_{ZINB}(X_i = y_i)) = \begin{cases} \sum ln(\omega_i) + (1-\omega_i)(\frac{1}{1+\alpha\lambda_i})^{1/\alpha}, \text{ if } y_i = 0; \\ \sum ln(\omega_i) + ln\Gamma(\frac{1}{\alpha} + y_i) - ln\Gamma(1+y_i) - ln\Gamma(\frac{1}{\alpha}) \\ +\frac{1}{\alpha}ln(\frac{1}{1+\alpha\lambda_i}) + y_i ln(1 - \frac{1}{1+\alpha\lambda_i}), \text{ if } y_i > 0; \end{cases} \tag{5.21}$$

Poisson regression

When applying the Poisson regression, the general form of λ_{10Poi} is given by $e^{\sum_{j=1}^m \beta_0+\beta_j x_j}$. Therefore, we need to transform Eq. (5.4) into the following expression:

$$\lambda_{10Poi} = \begin{cases} 0, \text{ if } F_{RAcc} = 0,\ V = 0 \text{ or } T = 0; \\ exp(K_1 + C_F \times F_{RAcc} + C_{CM} \times CM + C_{Profile} \times I_{Profile} + C_{Align} \times I_{Align} + \\ C_{Wid} \times Wid + C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg}), \text{ if } F_{RAcc} \neq 0, \\ V \neq 0, \text{ and } T \neq 0; \end{cases} \tag{5.22}$$

Table 5.4. Regression results of λ_{10Poi}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_1	-9.562	0.440	-21.714	×
F_{RAcc}	C_F	0.636	0.332	1.913	
CM	C_{CM}	0.005 (6)	2.949e-04	17.144	×
$I_{Profile}$	$C_{Profile}$	-0.076	0.122	-0.621	
I_{Align}	C_{Align}	0.326 (2)	0.069	4.756	×
Wid	C_{Wid}	0.206 (3)	0.026	8.051	×
$Leng$	C_{Leng}	0.030 (4)	0.009	3.232	×
RSL	C_{RSL}	0.011 (5)	0.001	7.895	×
F_{Reg}	C_{Reg}	1.725 (1)	0.334	5.165	×

The results estimated through the Poisson regression approach are shown in Table 5.4. According to these results, being similar to the NLS case, one can notice that the road profile is not significant ($|t-statistic| = 0.621 << 1.96$). On the other hand, with an exponential form, the impact of road accident factor F_{RAcc} is weakened, namely the impact of F_{RAcc} with an exponential form is not significant when using Poisson regression approach ($|t-statistic| = 1.913 < 1.96$). Furthermore, according to the sensitive degrees of these parameters with the exponential form (rank indicated in brackets), once again the LX-accident-prone region factor is the most sensitive contributor among these parameters.

NB regression

When applying the NB regression, the general form of λ_{10NB} is given by $e^{\sum_{j=1}^{m}\beta_0+\beta_j x_j+\varepsilon}$, and it still requires to be expressed by Eq. (5.22). The dispersion parameter α is estimated at 3.2394 in our study through the iterative estimation algorithm automatically. The estimated results of the NB regression are shown in Table 5.5.

Table 5.5. Regression results of λ_{10NB}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_1	-9.424	0.457	-20.615	×
F_{RAcc}	C_F	0.616	0.343	1.793	
CM	C_{CM}	0.006 (6)	3.762e-04	16.493	×
$I_{Profile}$	$C_{Profile}$	-0.107	0.126	-0.850	
I_{Align}	C_{Align}	0.298 (2)	0.072	4.159	×
Wid	C_{Wid}	0.199 (3)	0.028	7.173	×
$Leng$	C_{Leng}	0.031 (4)	0.010	3.201	×
RSL	C_{RSL}	0.010 (5)	0.001	7.034	×
F_{Reg}	C_{Reg}	1.508 (1)	0.351	4.294	×

According to the results associated with the NB regression approach, it is worth noticing that the road profile is still not significant ($|t - statistic| = 0.850 << 1.96$). One can also notice that the impact of F_{RAcc} with an exponential form is not significant as well, when using the NB regression approach ($|t - statistic| = 1.793 < 1.96$). Moreover, according to the sensitive degrees of these parameters with the exponential form (rank indicated in brackets), the LX-accident-prone region factor is still the most sensitive contributor among these parameters.

ZIP regression

When applying the ZIP regression, the general form of λ_{10ZIP} is given by $e^{\sum_{j=1}^{m} \beta_0 + \beta_j x_j}$, and it still requires to be expressed by Eq. (5.22). The estimated results of the ZIP regression are shown in Tables 5.6 (for nonzero observations) and 5.7 (for zero-inflation observations).

Table 5.6. Count model regression results of λ_{10ZIP}

Parameter	Coefficient	Estimated value	Standard error	$t - statistic$	Significant
	K_1	-1.128e+01	7.586e-01	-14.867	×
F_{RAcc}	C_F	3.717e-01	4.202e-01	0.885	
CM	C_{CM}	6.221e-03 (4)	4.336e-04	14.347	×
$I_{Profile}$	$C_{Profile}$	-1.855e-01	1.513e-01	-1.226	
I_{Align}	C_{Align}	1.483e-01	8.786e-02	1.688	
Wid	C_{Wid}	4.397e-01 (2)	6.625e-02	6.636	×
$Leng$	C_{Leng}	3.971e-02	1.725e-02	1.904	
RSL	C_{RSL}	1.432e-02 (3)	2.069e-03	6.921	×
F_{Reg}	C_{Reg}	2.319 (1)	6.655e-01	3.484	×

Table 5.7. Zero-inflation model regression results of λ_{10ZIP}

Parameter	Coefficient	Estimated value	Standard error	$t - statistic$	Significant
	K_1	-1.574e+01	4.276	-3.680	×
F_{RAcc}	C_F	-1.104	1.646	-0.671	
CM	C_{CM}	1.584e-03	1.450e-03	1.093	
$I_{Profile}$	$C_{Profile}$	-4.355e-01	6.531e-01	0.505	
I_{Align}	C_{Align}	-1.185	6.141e-01	-1.931	
Wid	C_{Wid}	1.024 (2)	2.241e-01	4.571	×
$Leng$	C_{Leng}	8.231e-02	4.190e-02	1.964	
RSL	C_{RSL}	4.117e-02 (3)	1.449e-02	2.840	×
F_{Reg}	C_{Reg}	5.861 (1)	1.748	3.353	×

According to the results associated with the ZIP regression approach, it is worth noticing that, as for the nonzero related model, F_{RAcc}, $I_{Profile}$, I_{Align} and $Leng$ are not significant (< 1.96). Moreover, according to the sensitive degrees of other significant parameters with the exponential form (rank indicated in brackets), the LX-accident-prone region factor is still the most sensitive contributor among these parameters. While as for the zero-inflation model, only the Wid, RSL and F_{Reg} are significant (> 1.96).

ZINB regression

When applying the ZINB regression, the general form of λ_{10ZINB} is given by $e^{\sum_{j=1}^{m}\beta_0+\beta_j x_j+\varepsilon}$, and it still requires to be expressed by Eq. (5.22). The values of dispersion parameter α for nonzero observations and zero-inflation observations are estimated at 3.8102 and 1.4069 respectively in our study through the iterative estimation algorithm automatically. The estimated results of the ZINB regression are shown in Tables 5.8 (for nonzero observations) and 5.9 (for zero-inflation observations).

Table 5.8. Count model regression results of λ_{10ZINB}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_1	-7.128	0.734	-9.709	×
F_{RAcc}	C_F	0.671	0.413	1.624	
CM	C_{CM}	4.486e-03 (3)	4.991e-04	8.990	×
$I_{Profile}$	$C_{Profile}$	-5.886e-02	0.144	-0.406	
I_{Align}	C_{Align}	0.371 (1)	8.274e-02	4.495	×
Wid	C_{Wid}	0.145 (2)	4.558e-02	3.175	×
$Leng$	C_{Leng}	3.219e-03	1.203e-02	0.268	
RSL	C_{RSL}	2.558e-03	1.954e-03	1.309	
F_{Reg}	C_{Reg}	0.795	0.446	1.783	

Table 5.9. Zero-inflation model regression results of λ_{10ZINB}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_1	-4.036	2.190	-6.709	×
F_{RAcc}	C_F	0.260 (1)	1.456	2.179	×
CM	C_{CM}	6.685e-02 (2)	1.838e-02	3.636	×
$I_{Profile}$	$C_{Profile}$	0.705	0.544	1.296	
I_{Align}	C_{Align}	0.535	0.328	1.632	
Wid	C_{Wid}	8.873e-02	0.180	0.491	
$Leng$	C_{Leng}	0.114	6.639e-02	1.725	
RSL	C_{RSL}	5456e-03	6.629e-03	0.823	
F_{Reg}	C_{Reg}	1.632	1.679	0.972	

According to the results associated with the ZINB regression approach, it is worth noticing that, as for the nonzero related model, CM, I_{Align} and Wid are significant ($>$ 1.96). One can also notice that according to the sensitive degrees of the three parameters (rank indicated in brackets), the LX width is the most sensitive contributor among them. While as for the zero-inflation model, only the F_{RAcc} and CM are significant ($>$ 1.96).

5.4 Model quality validation and predictive accuracy assessment

In this section, we will assess the quality of our prediction models while determining an appropriate statistical distribution to be combined with the models, in such a way as to ensure the most accurate estimation of the probability of accidents occurring at a given SAL2 in a given year.

5.4.1 Model quality comparison between λ_{10P} and λ_{10Y}

In order to validate the quality of the different prediction model variants, the Monte-Carlo test for randomly sampling annual accident frequencies which meet the condition that the predicted annual accident frequency at a given SAL2 is equal to or more than the observed annual accident frequency at the SAL2, is firstly performed (considering a safety strict principle, the predicted annual accident frequency should not be lower than the observed annual accident frequency). Then, the percentages of randomly sampled annual accident frequencies that meet this condition are computed to compare with the actual percentages of specified entire sampled annual accident frequencies (e.g., as for the entire 80,000 annual accident frequencies sampled out of 83,320, the actual entire percentage is computed as

80,000/83,320; while k annual accident frequencies within the 80,000 frequencies meet the above condition. Thus, the percentage of randomly sampled annual accident frequencies meeting this condition is computed as k/83,320).

Table 5.10 shows the Monte-Carlo test results. One can notice that, for the specified entire random sampling size 80,000, 40,000, 10,000, 5,000 and 500, the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition computed using λ_{10Y} are all closer to the actual percentages of specified entire sampled annual accident frequencies, compared with the percentages of randomly sampled annual accident frequencies computed using λ_{10P}. Moreover, the similarity between the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition, which are computed using λ_{10Y}, and the actual specified entire percentages is relatively high.

Although the Monte-Carlo test results indicate that the λ_{10Y} model seems more appropriate, the tested percentages of annual accident frequencies sampled according to the aforementioned condition closer to the actual percentages are not able to thoroughly attest to the fact that the quality of λ_{10Y} model is definitely better, since the predicted accident frequency may be much higher than the accident frequency observed. Therefore, further statistical tests are required to comprehensively evaluate the model quality.

Table 5.10. Monte-Carlo test results for λ_{10Y} and λ_{10P}

# Samples	Actual percentage of annual accident frequencies sampled	λ_{10Y}-model estimated percentage of annual accident frequencies sampled	λ_{10P}-model estimated percentage of annual accident frequencies sampled
80,000	0.96015	0.95482	0.94191
40,000	0.48008	0.47747	0.45463
10,000	0.12002	0.11946	0.11416
5,000	0.06001	0.05959	0.05665
500	0.00600	0.00598	0.00576

Akaike's information criterion (AIC) [Bozdogan 1987], the Bayesian information criterion (BIC) [Weakliem 1999], the Pearson chi-square statistic (PCS) test [Pearson 1900] and the degree of freedom (DF) are used to evaluate the goodness of fit (GOF) of the model. They can be respectively expressed as follows:

$$\mathrm{AIC} = n + n \times ln(2\pi) + n \times ln(\mathrm{RSS}/n) + 2(l+1) \tag{5.23}$$

$$\mathrm{BIC} = n + n \times ln(2\pi) + n \times ln(\mathrm{RSS}/n) + (l+1)ln(n) \tag{5.24}$$

$$\mathrm{PCS} = \sum_{i=1}^{n} \frac{(O_i - \lambda_i)^2}{\lambda_i} \tag{5.25}$$

$$\mathrm{DF} = n - (l+1) \tag{5.26}$$

where n is the sample size; RSS is the sum of the squares of residuals between the annual accident frequencies observed and the annual accident frequencies estimated; l is the number of independent exponential parameters; O_i is the annual accident frequency observed and λ_i is the annual accident frequency expected.

The AIC and BIC are two statistical measures to test the relative quality of models for a given set of data. Smaller AIC and BIC values indicate a better model fitting. The PCS test is used to determine if there is a significant difference between the values expected and the values observed. The PCS is roughly equal to DF if the model fits the data perfectly without any dispersion. In other words, the closer the PCS is to the DF, the better the model fits the data [Lu 2016]. These statistical test results are shown in Table 5.11 with the goodness ranked in brackets. Some findings can be noticed: 1) considering AIC and BIC, the λ_{10Y} model gives better results, since the AIC and BIC values corresponding to the λ_{10Y} model are much smaller than those for the λ_{10P} model; 2) as for PCS, the λ_{10Y} model is also the preferred one, since the PCS of the λ_{10Y} model is closer to DF (DFs of the λ_{10Y} and the λ_{10P} are considerably approximative).

5.4.2 Model quality comparison among variants of λ_{10Y}

Table 5.12 shows the Monte-Carlo test results. One can notice that, for the specified entire random sampling size 80,000, 40,000, 10,000, 5000 and 500, the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition computed using λ_{10Y} are all closer to the actual percentages of specified entire sampled annual accident frequencies, compared with these computed using λ_{10Poi}, λ_{10NB}, λ_{10ZIP} and λ_{10ZINB}, respectively. Moreover, the similarity between the percentages of randomly sampled annual accident frequencies meeting the aforementioned condition, which are computed using λ_{10Y}, and the actual specified entire percentages is relatively high.

Although the Monte-Carlo test results indicate that the λ_{10Y} model is more appropriate, the tested percentages of annual accident frequencies sampled according to the aforementioned condition do not allow for thoroughly attesting that the quality of λ_{10Y} model is definitely better, since the predicted accident frequency may be much higher

Table 5.11. Model GOF comparison between λ_{10P} and λ_{10Y}

Parameter	$\boldsymbol{\lambda_{10Y}}$	$\boldsymbol{\lambda_{10P}}$
Railway traffic characteristics		
Average daily railway traffic	×	×
Railway speed limit	×	
Roadway traffic characteristics		
Average daily road traffic	×	×
Annual road accidents	×	×
LX characteristics		
Alignment	×	
Profile	×	
LX width	×	
Crossing length	×	
Region	×	
AIC	-190,744 (1)	-190,591 (2)
BIC	-190,670 (1)	-190,573 (2)
PCS	65,796 (1)	53,108 (2)
DF	83,313	83,319
Goodness score (the lower, the better)	3	6

than the accident frequency observed. Therefore, the Monte-Carlo test should be seen as a preliminary check and further statistical tests are required to comprehensively evaluate the model quality.

Table 5.12. Monte-Carlo test results for variants of λ_{10Y}

# Samples	Actual percentage of annual accident frequencies sampled	λ_{10Y}-estimated percentage of annual accident frequencies sampled	λ_{10Poi}-estimated percentage of annual accident frequencies sampled	λ_{10NB}-estimated percentage of annual accident frequencies sampled	λ_{10ZIP}-estimated percentage of annual accident frequencies sampled	λ_{10ZINB}-estimated percentage of annual accident frequencies sampled
80,000	0.96015	0.95482	0.94167	0.94307	0.94511	0.94527
40,000	0.48008	0.47747	0.45667	0.46669	0.45701	0.47009
10,000	0.12002	0.11946	0.11414	0.11544	0.11514	0.11557
5000	0.06001	0.05959	0.05821	0.05881	0.05831	0.05876
500	0.00600	0.00598	0.00588	0.00588	0.00589	0.00590

The results of AIC, BIC and PCS statistical tests are shown in Table 5.13 with the goodness ranked in brackets. The following findings are obtained: 1) considering AIC and BIC, the λ_{10Y} model gives better results, since the AIC and BIC values corresponding to the λ_{10Y} model are much smaller than those for the λ_{10Poi}, λ_{10NB}, λ_{10ZIP} and λ_{10ZINB} models; 2) in terms of PCS test, the λ_{10Y} model is also the most effective one, since the PCS of λ_{10Y} model is closer to DF (DFs of λ_{10Y}, λ_{10Poi}, λ_{10NB}, λ_{10ZIP} and λ_{10ZINB} are considerably approximative).

5.4.3 Predictive accuracy assessment

As mentioned in section 5.4.1, the model quality of λ_{10Y} form is much better than that of λ_{10P} form. Therefore, in this section, we only assess the predictive accuracy in terms of accident occurrence according to the NLS regression model λ_{10Y} combined with the Poisson and NB distributions respectively, the Poisson regression model λ_{10Poi} naturally combined with the Poisson distribution, the NB regression model λ_{10NB} naturally combined with the NB distribution, the ZIP regression model λ_{10ZIP} naturally combined with the ZIP distribution and the ZINB regression model λ_{10ZINB} naturally combined with the ZINB distribution, so as to identify the combination that ensures the most effective prediction. One can refer to Appendix A [Liang 2018d] for the detailed predictive accuracy comparison between λ_{10Y} and λ_{10P}.

Here, the log-likelihood statistic test (LL) is adopted to assess the GOF of the accident frequency prediction model combined with a statistical distribution. The larger the LL, the more preferred the model [Lu 2016]. The mathematical expression of the LL is given as follows:

$$\mathrm{LL} = \sum_{i=1}^{n} ln(\hat{P}_i) \tag{5.27}$$

where n is the sample size and $\hat{P}_i$ is the estimated probability of accident frequency observed. $\hat{P}_i$ is computed respectively according to the accident frequency prediction model combined with the Poisson or the NB distribution.

Table 5.13. Model GOF comparison

Test	POI-λ_{10Y}	NB-λ_{10Y}	λ_{10Poi}	λ_{10NB}	λ_{10ZIP}	λ_{10ZINB}
AIC	-190,744 (1)	-190,744 (1)	-187,804 (5)	-189,942 (2)	-188,312 (4)	-189,826 (3)
BIC	-190,670 (1)	-190,670 (1)	-187,720 (5)	-189,858 (3)	-188,176 (4)	-189,935 (2)
PCS	65,796 (1)	65,796 (1)	125,495 (5)	123,715 (4)	118,185 (3)	110,496 (2)
DF	83,313	83,313	83,311	83,311	83,311	83,311
LL	-2,599 (2)	-2,596 (1)	-2,732 (6)	-2,711 (5)	-2,701 (4)	-2,631 (3)
Goodness score (the lower, the better)	5	4	21	14	15	10

LL test results are shown in Table 5.13. One can notice that, for the λ_{10Y} model combined with either the Poisson or NB distribution, its GOFs are significantly better than λ_{10Poi}, λ_{10NB}, λ_{10ZIP} and λ_{10ZINB} models' GOFs according to the LL test. Furthermore, the GOF of λ_{10Y} combined with the NB distribution (NB-λ_{10Y}) is better than when combined with the Poisson distribution (POI-λ_{10Y}).

Based on the predicted probability of the accident frequency observed, further Cumulative Distribution Function (CDF) analysis with regard to the Poisson and the NB distribution is performed to evaluate the quality of the accident frequency prediction model combined with the two statistical distributions. As shown in Fig. 5.4, the relationship between the CDF and the corresponding probability of a given event is obtained. $\hat{P}(\bullet)$ denotes the predicted probability of a given event obtained through the Poisson or NB distribution; O_i is the observed accident frequency and λ_i is the estimated accident frequency. The blue curve "CDF λ_{10NB}, $O_i > \lambda_i$" represents the CDF of the event " $O_i > \lambda_i$" obtained through the NB distribution combined with the λ_{10NB}; the red curve "CDF λ_{10NB}, $O_i <= \lambda_i$" represents the CDF of the event " $O_i <= \lambda_i$" obtained through the NB distribution combined with the λ_{10NB}; the green curve "CDF λ_{10Poi}, $O_i > \lambda_i$" represents the CDF of the event " $O_i > \lambda_i$" obtained through the Poisson distribution combined with the λ_{10Poi}; the violet curve "CDF λ_{10Poi}, $O_i <= \lambda_i$" represents the CDF of the event " $O_i <= \lambda_i$" obtained through the Poisson distribution combined with the λ_{10Poi}. The interpretation of the remaining curves involving the λ_{10Y} can be similarly obtained. Since some curves are almost covered by some others in Fig. 5.4, the main results of CDF analysis are extracted and shown in Table 5.14 for the sake of clarity.

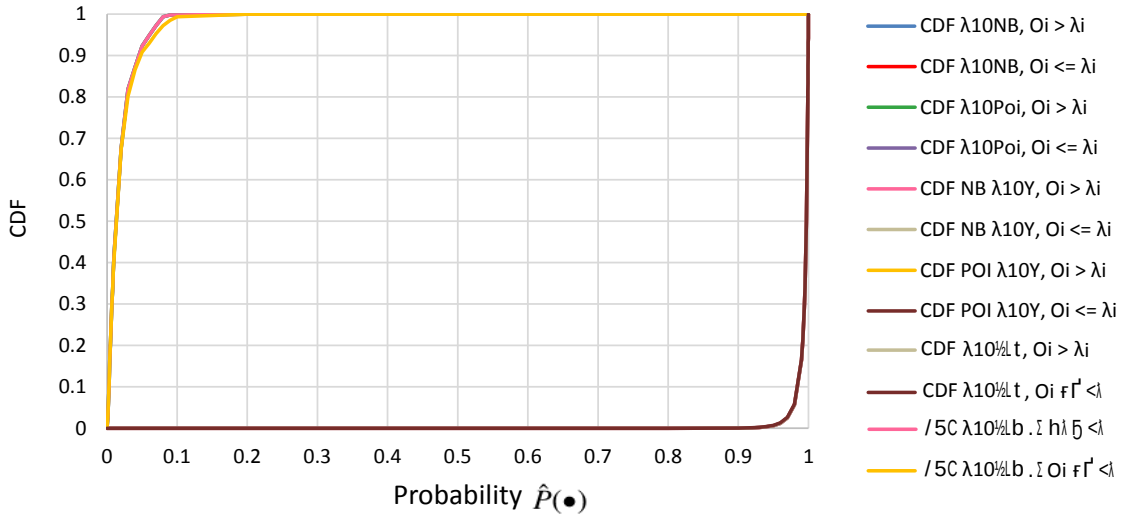


Fig. 5.4. CDF of the Poisson, NB, ZIP and ZINB distributions combined with the λ_{10Y}, λ_{10Poi}, λ_{10NB}, λ_{10ZIP} and λ_{10ZINB} models according to the estimated probability

Table 5.14. The extracted results of CDF analysis

Model CDF	$\hat{P}(O_i > \lambda_i) > 0.005$ (CDF in percent)	$\hat{P}(O_i > \lambda_i) > 0.05$ (CDF in percent)	$\hat{P}(O_i <= \lambda_i) > 0.95$ (CDF in percent)	$\hat{P}(O_i <= \lambda_i) > 0.995$ (CDF in percent)	Goodness score (the lower, the better)
CDF POI λ_{10Y}	78.89 (1)	9.17 (3)	99.27 (2)	65.94 (2)	8
CDF NB λ_{10Y}	79.10 (2)	7.68 (1)	99.36 (1)	66.07 (1)	5
CDF λ_{10Poi}	79.10 (2)	7.68 (1)	99.36 (1)	66.07 (1)	5
CDF λ_{10NB}	79.10 (2)	7.68 (1)	99.36 (1)	66.07 (1)	5
CDF λ_{10ZIP}	79.12 (3)	7.72 (2)	99.21 (3)	66.07 (1)	9
CDF λ_{10ZINB}	79.12 (3)	7.72 (2)	99.27 (2)	66.07 (1)	8

Table 5.14 indicates that:

1) CDF POI λ_{10Y}, $O_i > \lambda_i$:
In 78.89% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 9.17% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

2) CDF NB λ_{10Y}, $O_i > \lambda_i$:
In 79.10% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.68% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

3) CDF λ_{10Poi}, $O_i > \lambda_i$:
In 79.10% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.68% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

4) CDF λ_{10NB}, $O_i > \lambda_i$:
In 79.10% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.68% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

5) CDF λ_{10ZIP}, $O_i > \lambda_i$:
In 79.12% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.72% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

6) CDF λ_{10ZINB}, $O_i > \lambda_i$:
In 79.12% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.72% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

7) CDF POI λ_{10Y}, $O_i <= \lambda_i$:
In 99.27% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 65.94% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

8) CDF NB λ_{10Y}, $O_i <= \lambda_i$:
In 99.36% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

9) CDF λ_{10Poi}, $O_i <= \lambda_i$:
In 99.36% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

10) CDF λ_{10NB}, $O_i <= \lambda_i$:
In 99.36% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

11) CDF λ_{10ZIP}, $O_i <= \lambda_i$:
In 99.21% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

12) CDF λ_{10ZINB}, $O_i <= \lambda_i$:
In 99.27% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

Hence, according to the CDF analysis results shown in Table 5.14, in the cases of "$\hat{P}(O_i > \lambda_i) > 0.05$", "$\hat{P}(O_i <= \lambda_i) > 0.95$" and "$\hat{P}(O_i <= \lambda_i) > 0.995$", the λ_{10Y} combined with the NB distribution, the λ_{10Poi} combined with the Poisson distribution and the λ_{10NB} combined with the NB distribution show a better GOF value than that of the λ_{10Y} combined with the Poisson distribution. However, it is worth noticing that the distinction among the GOF values of these four types of model conjugations with regard to each case is not significant. Therefore, further analysis to assess the predictive accuracy of the three prediction models combined with the Poisson or NB distribution is carried out. As shown in Table 5.15, f_k denotes the percentage of samples of observed annual accident frequency with k accidents involved in a given year (f_k = the number of samples of observed annual accident frequency involving k accidents occurring in a given year / the total number of samples n). The estimated relative annual accident frequency reflected by estimated probabilities on average is computed as: $\hat{f}_k = \sum_{i=1}^{n} \hat{P}(X_i = k)/n$, where $\hat{P}(X_i = k)$ is the estimated probability of k accidents occurring at a given SAL2 in a given year. According to the goodness of predictive accuracy ranked in brackets, the λ_{10Y} model combined with the NB distribution shows an overall higher predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during the 10-year period, which means the probabilities of accident occurrence predicted by the λ_{10Y} model combined with the NB distribution are the closest to the actual probabilities of accident occurrence. In the case of 0 accident occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution takes the first place ($\hat{f}_k = 99.3915\%$). In the cases of 1, 2 and more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution takes the second place. Moreover, in the case of more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution shows a deviation of only 0.0002% compared with the actual f_k. In fact, there are no SAL2 LXs showing more than 2 accidents in the same year during the 10-year period considered.

Table 5.15. The predictive accuracy comparison between the Poisson, NB, ZIP and ZINB distributions

# Annual accidents considered (k)	Observed annual accident frequency (f_k in percent)	POI-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10Poi} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10NB} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10ZIP} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10ZINB} estimated relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.4371	99.3903 (3)	99.3915 (1)	99.3911 (2)	99.1871 (6)	99.3342 (4)	99.2851 (5)
1	0.5485	0.6033 (3)	0.5999 (2)	0.5905 (1)	0.7929 (6)	0.6005 (4)	0.6829 (5)
2	0.0144	0.0062 (5)	0.0077 (2)	0.0068 (4)	0.0188 (1)	0.0069 (3)	0.0229 (6)
> 2	0	0.0001 (1)	0.0002 (2)	0.0002 (2)	0.0012 (3)	0.0002 (2)	0.0015 (4)
Goodness score (the lower, the better)		12	7	9	16	13	20

Most recently, we obtained the updated incident/accident data up to 2017 from SNCF. Therefore, in order to validate the prediction performance of various variants of λ, a further analysis is carried out based on the data during the 10 years from 2008 to 2017, using the same method shown in Table 5.15. The updated results are shown in Table 5.16. According to the goodness of predictive accuracy ranked in brackets, the λ_{10Y} model combined with the NB distribution still shows the highest predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during this 10-year period. In the cases of 0, 1, 2 and more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution takes the first place in all cases. Moreover, in the case of more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution shows a deviation of only 0.0001% compared with the actual $f_k = 0$. In fact, there are no SAL2 LXs showing more than 2 accidents in the same year during this 10-year period considered either.

Based on the results of the various performed tests, there is no doubt that the λ_{10Y} model combined with the NB distribution is the most effective in terms of accident occurrence prediction among the four combinations considered. One can also notice that, although the NB distribution is more effective than the Poisson distribution when dealing with over-dispersed accident count, more extensive computations are required by NB model to estimate the model parameters as well as the dispersion parameter, so as to generate inferential statistics, compared with the Poisson model.

Table 5.16. The predictive accuracy comparison between the Poisson, NB, ZIP and ZINB distributions based on the data from 2008 to 2017

# Annual accidents considered (k)	Observed annual accident frequency (f_k in percent)	POI-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10Poi} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10NB} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10ZIP} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	λ_{10ZINB} estimated relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.6313	99.4603 (3)	99.5817 (1)	99.4801 (2)	99.2801 (6)	99.4348 (4)	99.3802 (5)
1	0.3616	0.4031 (3)	0.3813 (1)	0.3905 (2)	0.5129 (6)	0.4857 (4)	0.5071 (5)
2	0.0071	0.0032 (6)	0.0087 (1)	0.0037 (5)	0.0088 (2)	0.0039 (4)	0.0094 (3)
> 2	0	0.0001 (1)	0.0001 (1)	0.0002 (2)	0.0008 (3)	0.0002 (2)	0.0010 (4)
Goodness score (the lower, the better)		13	4	11	17	14	17

5.5 A comparison between λ_{10Y} and two existing reference models

In this section, we compare the present model λ_{10Y} with other two models which are widely used in existing related works. As mentioned in section 5.1, the first widely used model is given in Eq. (5.1) [Austin 2002, Oh 2006, Lu 2016]. In our study, this model can be specified as follows:

$$\lambda_{TV} = exp(K_2 + C_V \times V + C_T \times T + C_F \times F_{RAcc} + C_{Profile} \times I_{Profile} + C_{Align} \times I_{Align} + C_{Wid} \times Wid + C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg}) \tag{5.28}$$

where the average daily road traffic V and the average daily railway traffic T are applied separately in exponential form.

The second model as shown in Eq. (5.2) [Miranda-Moreno 2005, Saccomanno 2004], is specified as Eq. (5.29) in our study.

$$\lambda_{Mon} = exp(K_3 + C_M \times ln(V \times T) + C_F \times F_{RAcc} + C_{Profile} \times I_{Profile} + C_{Align} \times I_{Align} + C_{Wid} \times Wid + C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg}) \tag{5.29}$$

where the conventional traffic moment $V \times T$ are applied.

It should be noted that the ZIP and ZINB models were also investigated for λ_{TV} and λ_{Mon} but resulted in no higher goodness-of-fit values and a quite small number of significant parameters compared with the Poisson and NB models and hence, were not reported in this section. The Poisson and NB regression results of the λ_{TV} and λ_{Mon} are shown in Tables 5.17, 5.18, 5.19 and 5.20, respectively. One can notice that the impacts of road profile and road accident are still not significant in the λ_{TV} and λ_{Mon}.

The Monte-Carlo test is applied on the λ_{TV} and λ_{Mon}. The test results are shown in Table 5.21. It is worth noticing that the results computed using the λ_{10Y} are all closer to the actual percentages of specified entire sampled accident frequencies, compared with the results computed by the λ_{TV} and λ_{Mon}, respectively.

On the other hand, the AIC, BIC, PCS and LL tests and observed/estimated accident frequency comparison are given in Tables 5.22 and 5.23. According to the quality test results discussed in section 5.4.3, the λ_{10Y} combined with the NB distribution (NB-λ_{10Y}) shows the best prediction performance among the four investigated combinations. Therefore, we will only compare the NB-λ_{10Y} with the λ_{TV} and λ_{Mon} combined with the Poisson and NB distributions respectively in the following content.

Table 5.17. Poisson Regression results of λ_{TV}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_2	-9.807	0.413	-22.223	×
V	C_V	1.098e-04 (7)	1.613e-05	6.811	×
T	C_T	8.777e-03 (6)	1.115e-03	7.869	×
F_{RAcc}	C_F	0.636	0.333	1.913	
$I_{Profile}$	$C_{Profile}$	-1.445e-01	1.209e-01	-1.195	
I_{Align}	C_{Align}	3.319e-01 (2)	6.747e-02	4.919	×
Wid	C_{Wid}	2.059e-01 (3)	2.483e-02	8.292	×
$Leng$	C_{Leng}	3.952e-02 (4)	7.868e-03	5.024	×
RSL	C_{RSL}	1.154e-02 (5)	1.487e-03	7.759	×
F_{Reg}	C_{Reg}	1.750 (1)	3.463e-01	5.053	×

Table 5.18. NB Regression results of λ_{TV}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_2	-9.882	4.531e-01	-21.810	×
V	C_V	1.155e-04 (7)	1.683e-05	6.861	×
T	C_T	9.152e-03 (6)	1.234e-03	7.416	×
F_{RAcc}	C_F	0.607	3.402e-01	1.784	
$I_{Profile}$	$C_{Profile}$	-1.532e-01	1.243e-01	-1.232	
I_{Align}	C_{Align}	3.240e-01 (2)	6.988e-02	4.636	×
Wid	C_{Wid}	2.212e-01 (3)	2.579e-02	8.575	×
$Leng$	C_{Leng}	3.895e-02 (4)	8.415e-03	4.629	×
RSL	C_{RSL}	1.160e-02 (5)	1.529e-03	7.589	×
F_{Reg}	C_{Reg}	1.739 (1)	3.575e-01	4.864	×

Table 5.19. Poisson Regression results of λ_{Mon}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_2	-11.816	4.540e-01	-26.023	×
$ln(V \times T)$	C_M	4.036e-01 (2)	2.776e-02	14.538	×
F_{RAcc}	C_F	6.359e-01	3.325e-01	1.913	
$I_{Profile}$	$C_{Profile}$	-6.279e-02	1.205e-01	-0.521	
I_{Align}	C_{Align}	2.875e-01 (3)	6.799e-02	4.228	×
Wid	C_{Wid}	1.185e-01 (4)	3.296e-02	3.596	×
$Leng$	C_{Leng}	2.213e-02 (5)	9.530e-03	2.322	×
RSL	C_{RSL}	8.811e-03 (6)	1.350e-03	6.527	×
F_{Reg}	C_{Reg}	1.446 (1)	3.358e-01	4.307	×

Table 5.20. NB Regression results of λ_{Mon}

Parameter	Coefficient	Estimated value	Standard error	$t-statistic$	Significant
	K_2	-11.850	4.628e-01	-26.603	×
$ln(V \times T)$	C_M	4.034e-01 (2)	2.822e-02	14.297	×
F_{RAcc}	C_F	6.368e-01	3.382e-01	1.883	
$I_{Profile}$	$C_{Profile}$	-7.103e-02	1.230e-01	-0.578	
I_{Align}	C_{Align}	2.848e-01 (3)	6.960e-02	4.092	×
Wid	C_{Wid}	1.214e-01 (4)	3.361e-02	3.612	×
$Leng$	C_{Leng}	2.204e-02 (5)	9.752e-03	2.260	×
RSL	C_{RSL}	8.892e-03 (6)	1.368e-03	6.500	×
F_{Reg}	C_{Reg}	1.480 (1)	3.428e-01	4.316	×

Table 5.21. Monte-Carlo test results of λ_{10Y}, λ_{TV} and λ_{Mon}

# Samples	Actual percentage of annual accident frequencies sampled	λ_{10Y}-estimated percentage of annual accident frequencies sampled	λ_{TVPoi}-estimated percentage of annual accident frequencies sampled	λ_{TVNB}-estimated percentage of annual accident frequencies sampled	λ_{MonPoi}-estimated percentage of annual accident frequencies sampled	λ_{MonNB}-estimated percentage of annual accident frequencies sampled
80,000	0.96015	0.95482	0.93977	0.94011	0.94201	0.94356
40,000	0.48008	0.47747	0.45607	0.46141	0.45780	0.46367
10,000	0.12002	0.11946	0.11004	0.11471	0.11314	0.11509
5,000	0.06001	0.05959	0.05776	0.05783	0.05804	0.05872
500	0.00600	0.00598	0.00581	0.00584	0.00584	0.00587

As shown in Table 5.22, the AIC, BIC and PCS results related to the λ_{10Y} model are better than those for the λ_{TV} and λ_{Mon} models. Moreover, as for the LL test, the NB-λ_{10Y} is still the most preferred one. Besides, as for the observed/estimated accident frequency comparison as shown in Table 5.23, the NB-λ_{10Y} takes the first place according to the total goodness score. That means the NB-λ_{10Y} has higher predictive accuracy in terms of LX accident frequency than that of the POI-λ_{TV}, NB-λ_{TV}, POI-λ_{Mon} and NB-λ_{Mon}.

Table 5.22. Model GOF comparison among λ_{10Y}, λ_{TV} and λ_{Mon}

Test	NB-λ_{10Y}	POI-λ_{TV}	NB-λ_{TV}	POI-λ_{Mon}	NB-λ_{Mon}
AIC	-190,744 (1)	-177,914 (5)	-179,842 (4)	-183,714 (3)	-186,532 (2)
BIC	-190,670 (1)	-177,610 (5)	-179,738 (4)	-183,587 (3)	-186,191 (2)
PCS	65,796 (1)	121,715 (5)	119,133 (4)	118,511 (3)	115,634 (2)
DF	83,313	83,310	83,310	83,311	83,311
LL	-2,596 (1)	-2,722 (5)	-2,703 (3)	-2,705 (4)	-2,683 (2)
Goodness score (the lower, the better)	4	20	15	13	8

Table 5.23. The predictive accuracy comparison between the λ_{10Y} and the λ_{TV} and λ_{Mon}

# Annual accidents considered (k)	Observed annual accident frequency (f_k in percent)	NB-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{TV} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{TV} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{Mon} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{Mon} estimated relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.4371	99.3915 (1)	99.2912 (4)	99.2876 (5)	99.3215 (3)	99.3373 (2)
1	0.5485	0.5999 (1)	0.6005 (3)	0.6029 (4)	0.6005 (3)	0.6001 (2)
2	0.0144	0.0077 (2)	0.0068 (5)	0.0071 (3)	0.0069 (4)	0.0108 (1)
> 2	0	0.0002 (1)	0.0003 (2)	0.0016 (4)	0.0002 (1)	0.0013 (3)
Goodness score (the lower, the better)		5	14	16	11	8

5.6 Discussions

In this section, we will discuss the impact of variables considered in the λ_{10Y} model on SAL2 accident frequency in the following content.

It should be recalled that the corrected traffic moment is more effective in estimating automobile-involved LX accidents frequency compared with the conventional traffic moment, single average daily railway traffic or single average daily road traffic. Moreover, the higher the combined exposure of railway and roadway traffic, the higher the likelihood of an accident occurring. It should also be noted that the average daily railway traffic with a power of 0.646 has a more decisive impact on the LX accident frequency than the average daily road traffic with a power of 0.354, since the higher the railway traffic frequency appearing at SAL2 LXs, the much higher the SAL2 accident risk.

According to the aforementioned analyses, the form of λ_{10Y} highlights the impact of road accident factor F_{RAcc} (cf. Tables 5.4, 5.5, 5.17, 5.18, 5.19 and 5.20, the impact of F_{RAcc} is neglected in λ_{10Poi}, λ_{10NB}, λ_{TV} and λ_{Mon} models). The impact of road accidents on the risk level was likely to be ignored in the previous studies related to LX safety analysis. The present study has clearly shown that the accidents at LXs are above all road accidents and they highly depend on the road safety level. Moreover, the road accident factor in λ_{10Y} model reflects the time-dependent feature, since road accidents have an annual variation (vary every year). Accordingly, the risk related to LX accidents has an annual variation as well.

Among the LX characteristics, according to the sensitive degrees of variables ranked in Table 5.3, the risk of LX accidents is most sensitive to the region LX-accident-prone factor, followed by the road alignment, the LX width and the crossing length, successively. We originally introduce the region LX-accident-prone factor (cf. Table 5.2) in this study to interpret the variation of LX accident statistics with regard to various regions. In many past studies, the impact of LX local region is neglected. In fact, the regional accident history varies from one region to another, which correspondingly has varying degrees of impact on the LX accident frequency in different regions. Therefore, the region LX-accident-prone factor should be considered as a main parameter when estimating LX accident frequency in a wide area.

5.7 Summary

In this chapter, we set up an advanced accident frequency prediction model which takes into account a variety of impacting factors. This model allows for predicting accident occurrence with a considerably high accuracy and has a more appropriate form compared

with the existing models pertaining to LX accident prediction (cf. Eq. (5.1) and (5.2)). Moreover, such a model can be used to rank risky LXs, thus to identify LXs that are expected to have high possibility of accident occurrence compared with other LXs. It should also be noted that although the developed prediction model is tailored to SAL2 LX accidents in France, the general formula form of the model, the parameters considered and the methodologies for setting up such a model and validating the model quality can be advantageously applied to different countries.

The main contributions of the study reported in this chapter can be exposed as follows:

1) The selection of the impacting factors is backed by a scientific process that allows for identifying the important parameters to be considered and excluding the redundant ones.

2) The region LX-accident-prone factor which can interpret the impact of regional accident history on LX accident frequency is originally introduced in this study. Moreover, the significant impact of road accidents, generally ignored in the past studies, is well considered in our study.

3) The corrected traffic moment, a more effective factor, is proposed in the current study to replace the conventional traffic moment, single average daily railway traffic or single average daily road traffic when interpreting the likelihood of automobile-involved LX collisions.

4) A thorough and comprehensive validation process has been implemented. Namely, various statistical approaches, i.e., the Monte-Carlo, AIC, BIC and PCS tests, are adopted to ensure that the quality of λ_{10Y} prediction model is statistically approbatory. On the other hand, the LL test, CDF analysis and predictive accuracy analysis are conjunctively employed to determine the most appropriate statistical distribution for predicting the probability of LX accident occurrence. In our study, we find out that the NB distribution combined with λ_{10Y} prediction model shows high prediction performance. The results obtained by the aforementioned means have allowed us to consider the developed model to be trustworthy and sound. Such a thorough validation process is rarely achieved in existing related works.

5) A comparison between the present λ_{10Y} model and other two existing reference models is performed, which offers a clear viewpoint that the λ_{10Y} model is more effective than the other two models when predicting LX accident frequency and ranking risky LXs. Namely, the model set up the present study shows better prediction performance than the two reference models considered.

Chapter 6

Bayesian network based framework for LX risk reasoning

Sommaire

Overview

Now that we analyzed the motorist behavior factors (refer to chapter 4) and the static factors (as explained in section 1.2, the dysfunction of the LX control system is not considered in the present study) considered in statistical accident prediction model (refer to Chapter 5), we seek for a comprehensive approach to predict accident occurrence/consequences and make cause diagnosis while considering the both kinds of influential factors altogether. Therefore, in this chapter, a Bayesian network (BN) based framework for causal reasoning related to risk analysis is proposed. It consists of a set of integrated stages, namely, risk scenario definition, real field data collection and processing, BN model establishment and model performance validation. In particular, causal structural constraints are introduced to the framework for the purpose of combining empirical knowledge with data-driven, thus to identify effective causalities and avoid inappropriate structural connections. Then, the proposed framework is applied to risk analysis of LX accidents in France. In detail, the BN risk model is established based on the real field data of LX accidents/incidents and the model performance is validated. Moreover, forward inference and reverse inference based

on the BN risk model are performed to predict LX accident occurrence and quantify the contribution degree of various impacting factors respectively, so as to identify the most influencing factors. Besides, influence strength and sensitivity analyses are further carried out to scrutinize the influence strength of various causal factors on the LX accident occurrence and determine which factors the LX accident occurrence is most sensitive to.

The work reported in this chapter corresponds to the publications on the journals "Transportation Research Procedia" [Liang 2017c], "International Journal of Injury Control and Safety Promotion" [Liang 2018e], the conferences "RSSRail 2017" [Liang 2017a], "CTS 2018" [Liang 2018a] and submitted journal paper on "IEEE Trans. on Intelligent Transportation Systems" [Liang 2018b].

This chapter is structured as follows: section 6.1 exposes a general introduction to causal modeling techniques. Section 6.2 recalls the theory of probability which underlies BNs. Section 6.3 elaborates each stage of the BNI-RR framework. Section 6.4 is dedicated to the application of the BNI-RR framework to French LX risk analysis, particularly establishing and validating the BN risk model. Section 6.5 analyzes and discusses the outcomes of the BN risk model. Finally, section 6.6 offers the summary of this chapter.

6.1 Introduction

As investigated in chapter 2, there are a number of modeling techniques for risk analysis. FTs [Ericson 1999] and ETs [Bearfield 2005] are common techniques used for logical representation of a system for the purpose of risk reasoning analysis. MCs [Malyshkina 2009], PNs [Ghazel 2014] and FL [Niittymaki 1998] are also popular techniques for performing risk analysis in recent years. However, these approaches are unable to identify causality effectively.

Compared with the aforementioned modeling techniques, here we recall that BNs offer following interesting characteristics: flexibility of modeling, strong modeling power, high computational efficiency and, most importantly, the outstanding advantages involving the conjunction of domain expertise and automatic structure/parameters learning, causality analysis based on both forward inference (deductive reasoning) and reverse inference (abductive reasoning) [Weber 2012], as well as further influence and sensitivity analysis. Therefore, in this chapter, a novel framework of Bayesian Network (BN) based Inference for Risk Reasoning (BNI-RR) is proposed to deal with the analysis of accident causes and consequences at LXs. Besides, this framework describes a general risk analysis process that is not limited to the LX context and can be applied to other domains. Specifically, the present study involves: 1) developing an effective modeling framework, BNI-RR, for risk reasoning, which includes a set of comprehensive stages, i.e., risk scenario definition, real field data collection and processing, BN model establishment and model performance validation; 2) introducing causal structural constraints to the BNI-RR framework for the purpose of combining empirical knowledge with automatic learning approaches, so as to avoid inappropriate structural connections and highlight main causes; 3) developing BN models and performing corresponding analysis for French LX risk scenario using the BNI-RR framework, thus to identify potential impacting factors that contribute most to LX accident occurrence. The underlying aim is to pave the way toward discovering practical design measures and improvement recommendations to prevent accidents at LXs.

6.2 Bayesian Network

A BN is a graphical model that can be characterized by its structure and a set of parameters known as conditional probability tables (CPTs) [Jensen 1996]. $BN = (P, \mathcal{G})$, where P represents the parameters of conditional probabilities assigned to the arcs, while $\mathcal{G}$ defines the model structure. In fact, $\mathcal{G} = (N, L)$ is a Directed Acyclic Graph (DAG) that comprised of a finite set of nodes (N) linked by directed arcs (L). The nodes represent random variables and directed arcs between pairs of nodes represent dependencies between the variables. For instance, a three-variable BN is shown in Fig. 6.1. This net shows a "V" structure while the conditional probabilities of states are defined in each node.

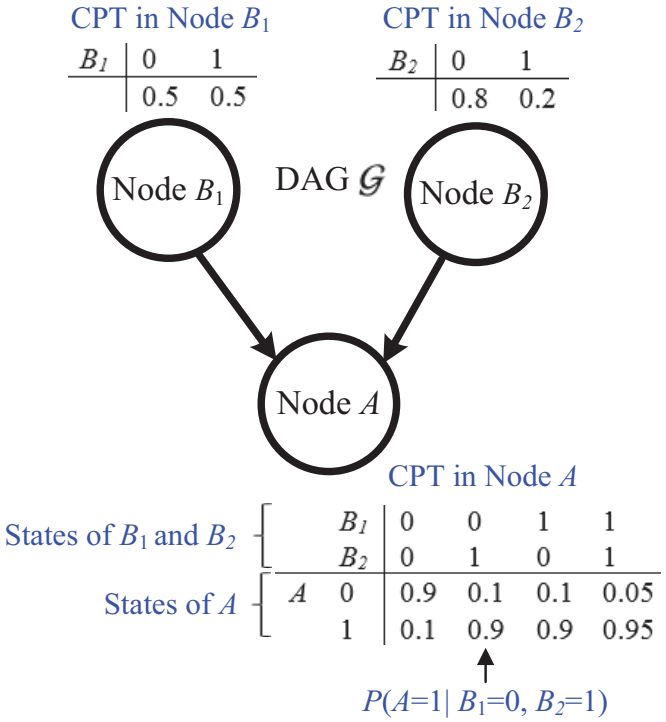


Fig. 6.1. An instance of a three-variable BN

The semantics of BN is based on the theory of probability. Assume that there is a set of mutually exclusive events: $B_1, B_2, \ldots, B_n$ and a given event A such that $P(A)$ can be expressed as follows:

$$P(A) = \sum_{i=1}^{n} P(B_i)\, P(A|B_i) \tag{6.1}$$

According to Bayes' formula:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)} \tag{6.2}$$

Hence, Eq. (6.2) can be converted into:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)} \tag{6.3}$$

Here, $P(B_i)$ is called the prior probability, while $P(B_i|A)$ is the posterior probability. Therefore, when the CPTs of variables are given in a BN, the posterior probability can be computed.

For any set of random variables in a BN, the joint distribution can be computed through conditional probabilities using the chain rule as shown in Eq. (6.4):

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v=1}^{n} P(X_v = x_v | X_{v+1} = x_{v+1}, \ldots, X_n = x_n) \tag{6.4}$$

Due to the conditional independence, X_v only relates to its parent node $Pa(X_v)$ and is independent of the other nodes. Hence, Eq. (6.4) can be rewritten as follows:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v=1}^{n} P(X_v = x_v | Pa(X_v)) \tag{6.5}$$

6.2.1 Modeling Tool

It is worthwhile to mention here that GeNIe [GeNIe 1999] is used as the BN modeling tool to build our BN risk model and perform analyses related the BN risk model. This tool offers interesting features in terms of graphical user interface and facility to operate and modify BN models. Moreover, it integrates extended functions for further analysis based on BN models.

Probability theory and decision theory supply tools for combining observations and optimizing decisions. The users of GeNIe will notice that this important premise is reflected in its functionality and, most importantly, its user interface. The principles of decision-analytic decision support, implemented in GeNIe can be applied in practical decision support systems (DSSs). In fact, quite a number of probabilistic decision support systems have been developed, in which GeNIe plays the role of a developer's environment and SMILE plays the role of the reasoning engine. A decision support system based on SMILE can be equipped with a dedicated user interface.

GeNIe supports the following node types: Chance nodes, drawn as ovals, denote uncertain variables.

Deterministic nodes, usually drawn as double-circles or double-ovals, represent either constant values or values that are algebraically determined from the states of their parents. In other words, if the values of their parents are known, then the value of a deterministic node is also known with certainty. Deterministic nodes are quantified similarly to Chance

nodes. The only difference is that their probability tables contain all zeros and ones (note that there is no uncertainty about the outcome of a deterministic node once all its parents are known).

Decision nodes, drawn as rectangles, denote variables that are under decision maker's control and are used to model decision maker's options. Decision nodes in GeNIe are always discrete and specified by a list of possible states/actions.

Value nodes (also called Utility nodes), drawn as hexagons, denote variables that contain information about the decision maker's goals and objectives. They express the decision maker's preferences over the outcomes over their direct predecessors.

Equation nodes, which are relatives of chance nodes, drawn as ovals with a wave symbol, denoting that they can take continuous values. Instead of a conditional probability distribution table, which describes the interaction of a discrete node with its parents, an equation node contains an equation that describes the interaction of the equation node with its parents. The equation can contain noise, which typically enters the equation in form of a probability distribution.

Deterministic equation nodes, drawn as double ovals with a wave symbol, denote equation nodes without noise, i.e., they are either constants or equations that do not contain a noise element. Once we know the states of their parents, the state of the child is, thus, determined.

Submodel nodes, drawn as rounded rectangles, denote submodels, i.e., conceptually related groups of variables. Submodel nodes are essentially holders for groups of nodes, existing only for the purpose of the user interface, and helping with making models manageable.

Arcs between nodes denote direct influences between them. One remark about editing diagrams is that GeNIe does not allow moving arcs between nodes, i.e., it is not possible to select and drag the head or the tail of an arc from one node to another. If this is what you want, the way to accomplish this task is to first delete the original arc and then create a new arc. These operations have serious consequences on the definitions of the nodes pointed by the heads of the arcs, namely deleting an arc deletes a portion of the definition of the node, adding an arc leads to a default extension of that definition. GeNIe tries to minimize the impact of adding and deleting arcs in terms of changing the conditional probability distributions. Whenever you add an arc, which amounts to adding a dimension to the child variable's conditional probability table, GeNIe will duplicate the current table, preserving the numbers from the original table. Whenever you delete an arc, which amounts to reducing a dimension of the child variable's conditional probability table, GeNIe will remove only a part of the table, preserving the rest.

6.2.2 Bayesian inference

BNs allow for performing Bayesian inference, i.e., computing the impact of observing values of a subset of the model variables on the probability distribution over the remaining variables. For example, observing a set of symptoms, captured as variables in a medical diagnostic model, allows for computing the probabilities of diseases captured by this model.

Bayesian inference, also referred to as Bayesian updating or belief updating, is based on the numerical parameters captured in the model. The structure of the model, i.e., an explicit statement of independences in the domain, helps in making the algorithms for Bayesian updating more efficient. All algorithms for Bayesian updating are based on a theorem proposed by Rev. Thomas Bayes (1702-1761) and known as Bayes theorem.

Belief inference in Bayesian networks is computationally complex. In the worst case, belief updating algorithms are NP-hard [Cooper 1990]. There exist several efficient algorithms, however, that make belief updating in graphs consisting of tens or hundreds of variables tractable. [Pearl 1986] developed a message-passing scheme that updates the probability distributions for each node in a Bayesian networks in response to observations of one or more variables. [Lauritzen 1988] and [Dawid 1992] proposed an efficient algorithm that first transforms a Bayesian network into a tree where each node in the tree corresponds to a subset of variables in the original graph. The algorithm then exploits several mathematical properties of this tree to perform probabilistic inference. Several approximate algorithms based on stochastic sampling have been developed. Of these, best known are probabilistic logic sampling [Lemmer 2014], likelihood sampling [Shachter 2013], backward sampling [Fung 1994], Adaptive Importance Sampling (AIS) [Cheng 1997], and quite likely the best stochastic sampling algorithm available at the moment, Evidence Pre-propagation Importance Sampling (EPIS) [Yuan 2002]. Approximate belief updating in Bayesian networks has been also shown to be worst-case NP-hard [Dagum 1993]. In most practical networks of the size of tens or hundreds of nodes, Bayesian updating is rapid and takes between a fraction of a second and a few seconds.

6.2.3 Decision support systems

Decision analysis is the art and practice of decision theory, an axiomatic theory prescribing how decisions should be made. Decision analysis is based on the premise that humans are reasonably capable of framing a decision problem, listing possible decision options, determining relevant factors, and quantifying uncertainty and preferences, but are rather weak in combining this information into a rational decision. Decision analysis comes with a set of empirically tested tools for framing decisions, structuring decision problems, quantifying uncertainty and preferences, discovering those factors in a decision model that are

critical for the decision, and computing the value of information that reduces uncertainty. While decision analysis is based on two quantitative theories, namely probability theory and decision theory, whose foundations are qualitative and based on axioms of rational choice. The purpose of decision analysis is to gain insight into a decision and not to obtain a recommendation.

Probabilistic decision support systems (DSSs), applicable to problems involving classification, prediction, and diagnosis, are a new generation of systems that are capable of modeling any real-world decision problem using theoretically sound and practically invaluable methods of probability theory and decision theory. Based on graphical representation of the problem structure, these systems allow for combining expert opinions with frequency data, gather, manage, and process information to arrive at intelligent solutions. Probabilistic DSSs are based on a philosophically different principle than rule-based expert systems. While the latter attempt to model the reasoning of a human expert, the former use an axiomatic theory to perform computation. The soundness of probability theory provides a clear advantage over rule-based systems that usually represent uncertainty in an ad-hoc manner, such as using certainty factors, leading to under-responsiveness or over-responsiveness to evidence and possibly incorrect conclusions. Probabilistic DSSs are applicable in many domains, among others in medicine (e.g., diagnosis, therapy planning), banking (e.g., credit authorization, fraud detection), insurance (e.g., risk analysis, fraud detection), military (e.g., target detection and prioritization, battle damage assessment, campaign planning), engineering (e.g., process control, machine and process diagnosis), and business (e.g., strategic planning, risk analysis, market analysis).

6.2.4 Probability

Decision theoretic and decision analytic methods quantify uncertainty by probability. It is quite important for a decision modeler to understand the meaning of probability. There are three fundamental interpretations of probability:

- Frequentist interpretation Probability of an event in this view is defined as the limiting frequency of occurrence of this event in an infinite number of trials. For example, the probability of heads in a single coin toss is the proportion of heads in an infinite number of coin tosses.

- Propensity interpretation Probability of an event in this view is determined by physical, objective properties of the object or the process generating the event. For example, the probability of heads in a single coin toss is determined by the physical properties of the coin, such as its flat symmetric shape and its two sides.

- Subjectivist interpretation The frequentist and the propensity views of probability are known as objectivist as they assume that the probability is an objective property of the physical world. In the subjectivist, also known as Bayesian interpretation, probability of an event is subjective to personal measure of the belief in that event occurring.

While the above three interpretations of probability are theoretical and subject to discussions and controversies in the domain of philosophy, they have serious implications on the practice of decision analysis. The first two views, known collectively as objectivist, are impractical for most real world decision problems. In the frequentist view, in order for a probability to be a meaningful measure of uncertainty, it is necessary that we deal with a process that is or at least can be imagined as repetitive in nature. While coin tosses provide such a process, uncertainty related to nuclear war is a rather hard case - there have been no nuclear wars in the past and even their repetition is rather hard to imagine. Obviously, for a sufficiently complex process, such as circumstances leading to a nuclear war, it is not easy to make an argument based on physical considerations. The subjectivist view gives us a tool for dealing with such problems and is the view embraced by decision analysis.

The subjectivist view interprets probability as a measure of personal belief. It is legitimate in this view to believe that the probability of heads in a single coin toss is 0.3, just as it is legitimate to believe that it is 0.5 as long as one does not violate the axioms of probability, such as one stating that the sum of probabilities of an event and its complement is equal to 1.0. It is also legitimate to put a measure of uncertainty on the event of nuclear war. Furthermore, this measure, a personal belief in the event, can vary among various individuals. While this sounds perhaps like a little too much freedom, this view comes with a rule for updating probability in light of new observations, known as Bayes theorem. There exist limits theorems that prove that if Bayes theorem is used for updating the degree of belief, this degree of belief will converge to the limiting frequency regardless of the actual value of the initial degree of belief (as long as it is not extreme in the sense of being exactly zero or exactly one). While these theorems give guarantees in the infinity, a reasonable prior belief will lead to a much faster convergence.

The subjectivist view makes it natural to combine frequency data with expert judgment. Numerical probabilities can be extracted from databases, can be based on expert judgment, or a combination of both. Obtaining numbers for probabilistic and decision-theoretic models is not really difficult. The process of measuring the degree of belief is referred to as a probability assessment. Various decision-analytic methods are available for probability assessment.

6.2.5 Discrete and continuous variables

One of the most fundamental properties of variables is their domain, i.e., the set of values that they can assume. While there is an infinite number of possible domains, they can be divided into two basic classes: discrete and continuous.

Discrete variables describe a finite set of conditions and take values from a finite, usually small, set of states. An example of a discrete variable is Success of the venture, defined in the tutorial on Bayesian networks. This variable can take two values: Success and Failure. Another example might be a variable Hepatitis-B, assuming values True and False. Yet another is Financial gain assuming three values: \$10K, \$20K, and \$50K.

Continuous variables can assume an infinite number of values. An example of a continuous variable is Body temperature, assuming any value between 30 and 45 degrees Celsius. Another might be Financial gain, assuming any monetary value between zero and \$50K.

Most algorithms for Bayesian networks are designed for discrete variables. To take advantage of these algorithms, most Bayesian network and influence diagram models include discrete variables or conceptually continuous variables that have been discretized for the purpose of reasoning. While the distinction between discrete and continuous variables is clear, the distinction between discrete and continuous quantities is rather vague. Many quantities can be represented as both discrete and continuous. Discrete variables are usually convenient approximations of real world quantities, sufficient for the purpose of reasoning. And so, success of a venture might be represented by a continuous variable expressing the financial gain or stock price, but it can also be discretized to [Good, Moderate, Bad] or to [\$5, \$20, \$50] price per share. Body temperature might be continuous but can be also discretized as Low, Normal, Fever, High fever. Experience in decision analytic modeling has taught that representing continuous variables by their three to five point discrete approximations perform very well in most cases.

6.2.6 BN learning

Parameter learning

Parameter learning is the process of using data to learn the distributions of a BN. GeNIe can estimate parameters of BNs using Expectation Maximization (EM) algorithm perform Maximum Likelihood (ML) [Dempster 1977] for a given structure. In the most general parameterization, when the data are fully observed, the ML estimation problem decomposes into independent sub-problems associated with each CPT.

Structural learning

Structural learning is the process of using data to learn the links of a BN. There are six classical structure learning approaches that are widely used to build BN structures:

1) The Bayesian Search (BS) algorithm is one of the earliest and the most popular algorithms used. It was introduced in [Cooper 2007] and later was refined in [Heckerman 1995]. It follows essentially a hill climbing procedure, generally guided by a scoring heuristic, with random restarts.

2) The Essential Graph Search (EGS) algorithm, proposed in [Dash 1999], performs a search for essential graphs based on a combination of the constraint-based search and BS approach.

3) The Greedy Thick Thinning (GTT) algorithm performs based on the BS approach and has been described in [Cheng 1997]. GTT starts with an empty graph and repeatedly adds the arc (without creating a cycle) that maximally increases the marginal likelihood until no arc addition results in a positive increase (thickening phase). Then, it repeatedly removes arcs until no arc deletion results in a positive increase in the marginal likelihood (thinning phase).

4) The Naïve Bayes approach [Good 1966] creates a Bayesian network including its structure and parameters, directly from data. In fact, the structure of a Naïve Bayes network is not learned but rather fixed by an assumption: the class variable is the only parent of all remaining feature variables and there are no other connections between the nodes of the network. Note that the Naïve Bayes structure assumes that the feature variables are independent conditional on the class parent variable, which leads to inaccuracies when they are not independent in reality.

5) The Augmented Naïve Bayes (ANB) algorithm is a semi-naive structure learning method based on the BS approach [Friedman 1997]. The ANB algorithm starts with a Naïve Bayes structure (i.e., the class variable is the only parent of all remaining feature variables) and adds connections between the feature variables to account for possible dependence between them and conditional on the class variable. There is no limit on the number of additional connections between the feature variables. The ANB algorithm is simple and has been found to perform reliably better than Naïve Bayes.

6) The Tree Augmented Naïve Bayes (TAN) algorithm is also a semi-naive structure learning method based on the BS approach [Friedman 1997]. Compared with The ANB algorithm, The TAN algorithm imposes the limit of only one additional parent of each feature variable (additional to the initial class variable that is the parent of every feature variable). The TAN algorithm is simple and performs better than Naïve Bayes as well.

6.2.7 Strength of influence

In fact, interactions between pairs of variables, denoted by directed arcs, may have different strength. It is often of interest to the modeler to visualize the strength of these interactions. GeNIe offers a functionality that pictures the strength of interactions by means of arc thickness. This is especially useful in the model building and testing phase. Model builders or experts can verify whether the thickness of arrows corresponds to their intuition. If not, this offers an opportunity to modify the parameters accordingly.

The strength of influence can be visualized by automatically varying the thickness of the arc connecting the nodes through GeNIe. It is quite intuitive to draw a thicker arc when the influence is strong. Thickness of arcs can be based on one of the three: Average (default), Maximum, and Weighted. Maximum uses the largest distance between distributions, Average takes the plain average over distances, and Weighted weighs the distances by the marginal probability of the parent node. There are two modes of strength of influences, namely the normalized and non-normalized mode. In normalized mode, the thickest possible arc is given to that arc that has the highest strength of influence. The thicknesses of all other arcs are calculated proportionally to the thickest arc. In the non-normalized mode (default), thickness is based on the absolute value of the distance.

6.2.8 Sensitivity analysis

Sensitivity analysis [Castillo 1997] is technique that can help validate the probability parameters of a Bayesian network. This is done by investigating the effect of small changes in numerical parameters (i.e., probabilities) on the output parameters (e.g., posterior probabilities). Highly sensitive parameters affect the reasoning results more significantly. Identifying them allows for a directed allocation of effort in order to obtain accurate results of a Bayesian network model.

GeNIe implements an algorithm proposed by [Kjærulff 2000] that performs simple sensitivity analysis in Bayesian networks. Roughly speaking, given a set of target nodes, the algorithm calculates efficiently a complete set of derivatives of the posterior probability distributions over the target nodes over each of the numerical parameters of the Bayesian network. These derivatives give an indication of importance of precision of network numerical parameters for calculating the posterior probabilities of the targets. If the derivative is large for a parameter P, then a small deviation in P may lead to a large difference in the posteriors of the targets. If the derivative is small, then even large deviations in the parameter make little difference in the posteriors.

6.3 BNI-RR framework

A modeling paradigm should view an influential network not merely as passive parsimonious codes for storing factual knowledge, but also as a computational architecture for reasoning about the knowledge. It means that the links in the network should be treated as the only pathways and activation units that direct and propel the flow of data in the process of querying and updating causal knowledge. In this section, while having in mind this principle, the BNI-RR framework is proposed. The BNI-RR framework can be illustrated as shown in Fig. 6.2. Namely, the process of BNI-RR approach consists of the following stages:

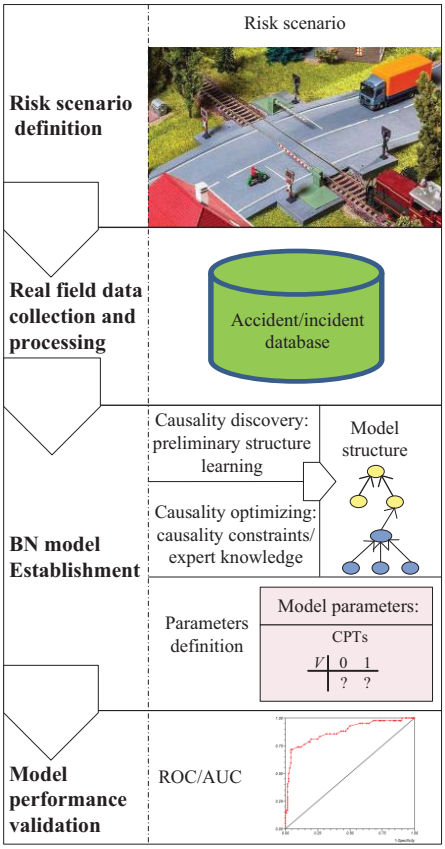


Fig. 6.2. The BNI-RR framework

1) Risk scenario definition: before performing risk analysis and in order to set the research target, a clear definition of the risk scenario boundary must be achieved. One should focus on this defined risk scenario to ensure that the follow-up study does not deviate from the original intention.

2) Real field data collection and processing: for risk analysis and cause diagnosis, real

field data related to the defined risk scenario need to be collected. These data should be recorded in a workable database and used as the basis of data processing. Data processing includes data merging/cleansing and data discretization, which is the basis of parameters learning and CPT definition.

3) BN model establishment: on the one hand, the model structure is constructed with regard to the combination of automatic structure learning and causality constraints derived from expert knowledge (cf. 6.3.1). On the other hand, the model CPTs are generated based on the post-processing field data. Model structure constructing in this stage will be elaborated in section 6.3.1.

4) Model validation: the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) [Hanley 1982] are adopted to validate the model performance of prediction. The ROC curve is a two-dimensional graph that can be obtained by plotting the true positive rate (TPR) (Y-axis) against the false positive rate (FPR) (X-axis) at various threshold settings [Powers 2011]. The TPR is known as the *sensitivity*, the *recall* or the probability of detection in machine learning. The FPR is known as the *fall-out* or the probability of false alarm. The ROC curve thus depicts relative trade-offs between benefits (true positives) and costs (false positives). In order to facilitate the evaluation of classifier performance, one may want to reduce ROC performance to a single scalar value that can represent the expected performance. A common method is to calculate the AUC which is a portion of the area of a unit square, and the value of which falls into the interval between 0 and 1. When using normalized units, the AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The ROC curve of a finite set of samples is based on a step function, and its AUC can be computed by the normalized Wilcoxon-Mann-Whitney (WMW) statistic [Yan 2003]:

$$AUC = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} I(x_i, y_j)}{m \times n} \tag{6.6}$$

where x_i, $i = 1, \ldots, m$, is the sample of positive classifier outputs; y_j, $j = 1, \ldots, n$, is the sample of negative classifier outputs and

$$I(x_i, y_j) = \begin{cases} 1, x_i > y_j \\ 0, \text{otherwise} \end{cases} \tag{6.7}$$

is based on pairwise comparisons between x_i and y_j.

6.3.1 BN model structure constructing

6.3.1.1 Causality discovery

As shown in Fig. 6.3, causality is the relationship between a cause and a consequence. Identifying such causal relationships is a crucial issue in the process of risk reasoning. In particular, a functional intelligent decision/prediction model should have the ability of making reasoning based on causal knowledge.

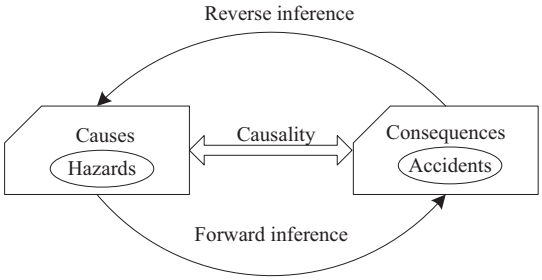


Fig. 6.3. Reasoning between causes and consequences

For instance, in railways, potential hazards such as human errors and environment aspects, may lead to incidents/accidents. Taking human errors for example, this can be expressed as a rule IF human errors occur, $THEN$ accidents may occur. Therefore, the DAG $\mathcal{G}_C$ of a causal network can be interpreted by a causal semantics as follows:

$$\mathcal{G}_C = \{IF, THEN, CAK\} \tag{6.8}$$

where:

$\mathcal{G}_C$ is a 3-tuple causal DAG;

IF is a set of causes, $IF = \{x_1, x_2, \ldots, x_n\}$;

$THEN$ is a set of consequences caused by the causes in IF, $THEN = \{y_1, y_2, \ldots, y_m\}$;

CAK represents the CAusal Knowledge, which is a set of directed pairs of the cause $x_i \in IF$ and the corresponding consequence $y_j \in THEN$: $CAK = \{(x_i, y_j) | x_i \in IF, i = 1, 2, \ldots, n; y_j \in THEN, j = 1, 2, \ldots, m\}$ while note that (x_i, y_j) is a directed variable pair that defines the structure of $\mathcal{G}_C$: $x_i \rightarrow y_j$ and cannot be reversed, which reflects the causal relationship between x_i and y_j at the same time. For example, in Fig. 6.1, $\mathcal{G}_C = \{IF = \{B_1, B_2\}, THEN = \{A\}, CAK = \{(B_1, A), (B_2, A)\}\}$.

Hence, by considering the causality in the BN, the states of the target variable can be predicted even when the states of the other factors are changed. More importantly, once a given state of the target variable is observed, the contribution of the impacting factors can be assessed. In practice, preliminary causality can be discovered through the six automatic structure learning mentioned in 6.2.6.

6.3.1.2 Causality optimizing

In terms of causal reasoning, one can notice that model structures learned based on the aforementioned approaches are often preliminary, even make no sense of reasonability. These preliminary structures are inconsistent with the causal relationships in reality, and in some cases, some connections are more likely correlations rather than causalities in reality and impede identification of important causes. Causalities can be identified from correlations, however, causalities are not equal to correlations. Many previous methods cannot achieve this important issue; therefore, causality optimizing is indispensable to be performed based on causal constraints, for the purpose of finely distinguishing causalities from correlations.

Pearl and Verma have stated that an intelligent modeling system should have the competence of translating direct observations to cause-effect relationships. Moreover, expert knowledge is significant to distinguish causalities from correlations in terms of causation [Pearl 1995]. Therefore, causal structural constraints [De Campos 2007] (CSCs) generated from expert knowledge are adopted to achieve causality optimizing in the present study.

In general, there are 3 types of directed CSCs of BNs: Existence Constraint (EC), Forbidden Constraint (FC), and Potential Directed Constraint (PDC). For instance, given a BN $\mathcal{N}$ and two variables x and y of $\mathcal{N}$, based on the definition of CAK, an EC $(x, y)_e$ means that there must be a direct connection from x to y; an FC $(x, y)_f$ means that there must not be a direct connection from x to y; a PDC $(x, y)_p$ means that if there exists a direct connection between x and y, it should be from x to y, while from y to x is not allowed. Utilizing PDCs can control constraint granularity and be perspicuous to describe a contrary edge orientation to inappropriate automatic learning structure.

To sum up, adopting jointly the above directed CSCs can effectively perform combination of automatic structure learning from observational data and expert knowledge. The detailed advantages are three-fold: 1) identifying unknown but potentially valuable causalities, especially when samples are limited, 2) verifying the already known causalities, and 3) avoiding inappropriate connections to facilitate highlighting main causes.

6.4 Application

In this section, the BNI-RR framework is applied to the risk analysis of French LXs. The risk analysis is carried out based on the real field accident/incident data collected by SNCF. In the sequel, we will discuss the various steps of the framework.

6.4.1 Risk scenario definition

As stated in chapter 1, we need to recall that "as shown in Table 1.1, SAL2 (more than 10,000) is the most widely used type of LX in France according to the LX data recorded by SNCF. Moreover, the accident/incident records show that more than 4,000 accidents at SAL2 contributed most to the total number of accidents at LXs from 1974 to 2014. In addition, according to SNCF statistics, the accidents at SAL2 LXs can be considered as the most representative for LX accidents in general. Besides, being given the number of SAL2 LXs, dealing with this LX category constitutes a priority issue for SNCF. On the other hand, according to the previous statistical analysis, one can notice that the motorized vehicle is the main transport mode causing accidents at LXs [Liang 2017c, Liang 2018d]. Considering the train/motorized vehicle (train-MV) collisions, SAL2 LXs also have the most accidents from 1978 to 2013 [Liang 2017c]." Therefore, in what follows, we consider the risk scenario corresponding to the situation where the "motorized vehicles cross SAL2 LXs when trains are approaching".

6.4.2 Data collection and processing

The approach for causal inference in the present study is based on field-observational-data. For the main purpose of assessing risk level and diagnosing causes, real field accident/incident data related to the defined risk scenario need to be collected. This is an important preparatory stage that is required prior to the establishment of the BN model. It should be noted that, in terms of ethics approval, the data collected in the present study do not hold any personal or private aspects.

SNCF Réseau investigated and recorded various attributes of LX accidents/incidents, such as railway and roadway traffic characteristics, surrounding characteristics of LXs and then, provided two accident/incident databases to support our study. The first database (D1) records the accident/incident data that cover SAL2 LXs in mainland France from 1990 to 2013. From D1, the sub-dataset (SD1) including the data ranging in the decade from 2004 to 2013 is selected, which provides reliable and sufficient information about both LX accidents and static railway, roadway and LX characteristics (considered as permanent characteristics related to LXs). Namely, the selected LX inventory presents the

LX identification number, the LX accident timestamp, the railway line involved, the LX kilometer point, the average daily railway traffic, the average daily road traffic, the rail speed limit, the LX length and width, the profile and alignment of the entered road and geographic region involved. There are 8,332 public SAL2 LXs included in SD1.

According to the statistics, the majority of train-MV accidents at LXs are caused by motorist violations. Due to the lack of accident causes in SD1, causal reasoning analysis cannot be performed with regard to the static factors and motorist behavior. Therefore, we need to utilize another database which records detailed accident causes. Fortunately, the second database (D2) contains the information about SAL2 LX accidents during the period from 2010 to 2013, namely, the LX identification number, the railway line involved, fatalities, injuries, and accident causes (including static factors and inappropriate motorist behavior). Thus, using the LX ID and the railway line ID, data merging of these two databases is carried out to create a new database (ND) containing the LX accident information, static railway, roadway and LX characteristics, the number of fatalities and injuries, and accident causes related to static factors and motorist behavior. This combined database ND covers LX accidents during a period of 4 years from 2010 to 2013, which forms the basis of our present study.

The accident causes were classified into three levels: primary, secondary and third-level causes. The various causes considered in this study are shown in Table 6.1. It should be noted that corrected moment (CM) which is a secondary cause, is a variant of the conventional traffic moment (refer to 5.3.1). Moreover, data discretization is applied on continuous causal variables. Namely, the continuous causal variables, i.e., “Average Daily Road Traffic”, “Average Daily Railway Traffic”, “Railway Speed Limit”, “LX Width”, “Crossing Length” and “Corrected Moment”, are divided into 3 groups such that each group has the same number of samples, based on the expert judgment and for the purpose of avoiding over-size model. As for the “Region Risk” factors corresponding to 21 regions in mainland France, they are divided into 3 groups as well, ranked according to the risk level in descending order, and each group contains 7 region risk factors, based on the expert judgment. As for the finite discrete causal variables, i.e., “Alignment”, “Profile”, “Stall on LX”, “Zigzag Violation”, “Blocked on LX” and “Stop on LX”, we allocate an individual state to each value of the variable.

6.4.3 BN modeling

In the subsequent sections, we will go through the various steps of the BN model development.

Table 6.1. Accident causal factors

Primary causes (PriC)	Secondary causes (SC)	Third-level causes (TC)	Explanation
Static Factors	Corrected Moment (CM)	Average Daily Railway Traffic (T) Average Daily Road Traffic (V)	$CM = V^{0.354} \times T^{0.646}$;
	Railway Speed Limit		The maximum permission speed of train within the LX section;
	Alignment		Horizontal road alignment shape: "straight", "curve" or "S";
	Profile		Vertical road profile shape: "normal" or "hump or cavity";
	LX width		The entered road width;
	Crossing Length		The length of LX that road vehicles need to cross;
	Region Risk		Highlighting the general LX-accident-prone region: *The number of SAL2 accidents over the observation period in the region considered/The number of SAL2 LXs in the region considered*;
Inappropriate Motorist Behavior	Stall on LX	Blocked on LX	A vehicle is blocked on the SAL2 LX by the external environment;
		Stop on LX	A motorist intentionally stops the vehicle on the SAL2 LX;
	Zigzag Violation		A vehicle skirts the half barriers to cross the SAL2 LX;

6.4.3.1 Variable definition

Based on the combined database ND, the pre-processed data of causal variables aforementioned in section 6.4.2 are organized as input sources which will be used to generate the CPTs of our BN risk model. On the other hand, consequence variables, i.e., "Fatalities", "Severe Injuries" and "Minor Injuries", are defined respectively with two states according to the domain expertise and the coefficient of variation (StdDev/Mean) [Reed 2002] of the three variables. Table 6.2 shows the statistical characterization of the numerical variables considered.

Table 6.2. Statistical characterization of numerical variables

Variable	Mean	Min	Max	StdDev
Annual Accident	0.0057	0	2	0.0776
Corrected Moment	51.4744	1.2781	938.5449	61.1367
Average Daily Railway Traffic	26.0636	0.5000	330	30.2413
Average Daily Road Traffic	826.8022	0.5700	2.5570e+04	1.7810e+03
Railway Speed Limit	92.4599	5	160	42.3829
Length	9.6766	3	59	3.8671
Width	5.4504	2	24	1.3569
Region Risk	0.3487	0.1739	0.7747	0.1194
Fatalities	0.2511	0	4	0.3543
Severe Injuries	0.2134	0	5	0.4867
Minor Injuries	0.5546	0	39	1.6162

Besides, in our BN risk model, an additional variable corresponding to the consequence severity [50126 1999] is defined according to the number of fatalities and injuries in a given SAL2 accident, based on the accident/incident statistics and expert judgment. The definition of consequence severity pertaining to an SAL2 accident is illustrated in Table 6.3. Five levels of consequence severity are set according to the number of fatalities, severe injuries and minor injuries caused by the accident, respectively. The consequence severity increases progressively from level 1 to 5. Thus, a summary of the states corresponding to

Table 6.3. Definition of consequence severity classification

Consequence severity	Level 1	Level 2	Level 3	Level 4	Level 5
$0 =$ fatalities, $0 \leq$ severe injuries < 2, $0 \leq$ minor injuries < 3;	×	–	–	–	–
$0 =$ fatalities, $0 \leq$ severe injuries < 2, $3 \leq$ minor injuries;	–	×	–	–	–
$0 =$ fatalities, $2 \leq$ severe injuries, $0 \leq$ minor injuries < 3;	–	–	×	–	–
$0 =$ fatalities, $2 \leq$ severe injuries, $3 \leq$ minor injuries;	–	–	–	×	–
$0 <$ fatalities;	–	–	–	–	×

each node in the BN risk model is given in Table 6.4.

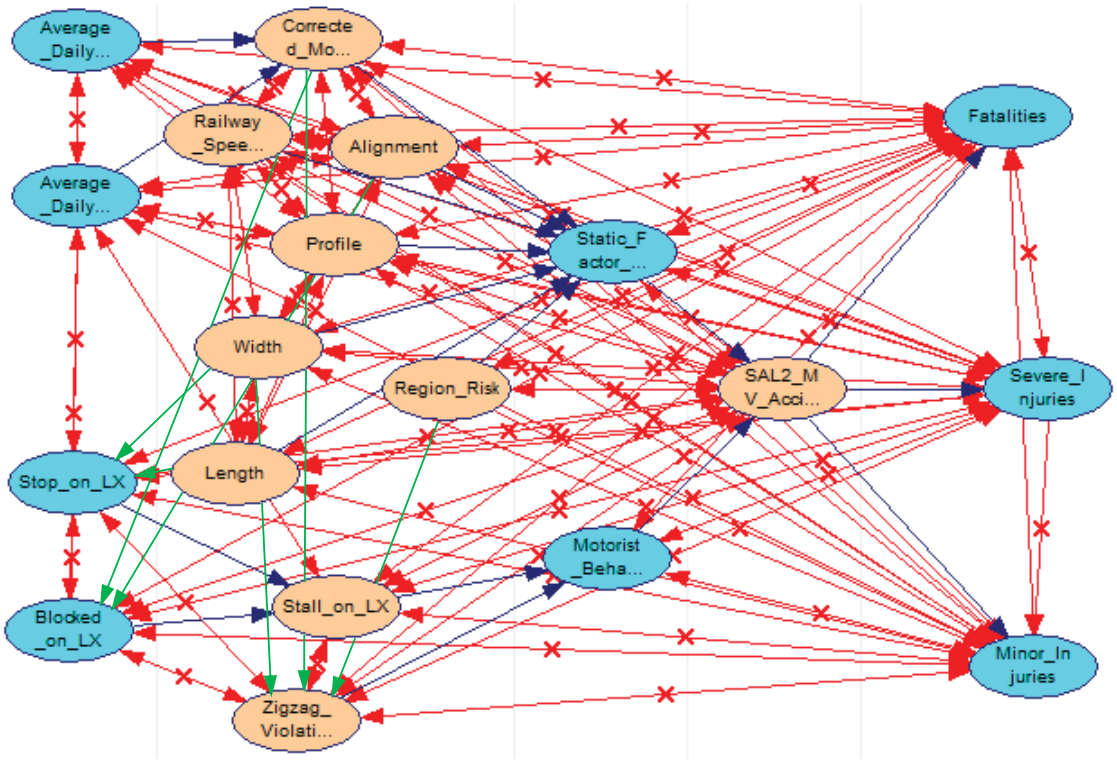


Fig. 6.4. CSCs identified for the BN risk model

6.4.3.2 Model structure establishment

In this stage, CSCs are adopted to set up our BN risk model. As shown in Table 6.1, the causal variables considered fall into two types: static factors (representing contextual information regarding LX characteristics, railway traffic characteristics, road traffic characteristics, cf. Table 5.2) and motorist behavior related factors. We firstly identify the internal CSCs within static factors and motorist behavior factors, respectively. Furthermore, it is worth noticing that there are some potential connections between static factors and motorist behavior. SNCF experts provide their knowledge about CSCs between these two types of factors, while checking the potential causal relationships. Therefore, the whole CSCs are identified as shown in Fig. 6.4. In this figure, blue, red and green arcs represent ECs, FCs and PDCs, respectively. Note that the "Consequence Severity" shown in Table 6.4 is a Deterministic node that is not considered in the process of CSC identification. In order to show these CSCs more clearly, we list them in Table 6.5. PDCs and some FCs are suggested by SNCF experts. With these CSCs, the final BN risk model is generated as shown in Fig. 6.8. In addition, CPTs are generated based on the post-processing real field accident/incident data.

Table 6.4. States of nodes in the BN risk model

Node name	Node property	Node state
TC nodes		
Average Daily Railway Traffic (ADRT)	Chance node	ADRT_below_9 ($0 \leq$ ADRT <9), ADRT_9_25 ($9 \leq$ ADRT <25), ADRT_25_up ($25 \leq$ ADRT);
Average Daily Road Vehicle (ADRV)	Chance node	ADRV_below_72 ($0 \leq$ ADRV <72), ADRV_72_403 ($72 \leq$ ADRV <403), ADRV_403_up ($403 \leq$ ADRV);
Blocked on LX (B)	Chance node	True, False;
Stop on LX (Stop)	Chance node	True, False;
SC nodes		
Corrected Moment (CM)	Chance node	CM_below_19 ($0 \leq$ CM <19), CM_19_49 ($19 \leq$ CM <49), CM_49_up ($49 \leq$ CM);
Railway Speed Limit (RLS)	Chance node	RLS_below_70 (0 km/h $\leq$ RLS <70 km/h), RLS_70_110 (70 km/h $\leq$ RLS <110 km/h), RLS_110_up (110 km/h $\leq$ RLS);
Alignment (A)	Chance node	Straight, C_shape, S_shape;
Profile (P)	Chance node	Normal, Hump_cavity;
Width (W)	Chance node	W_below_5 (0 m $\leq$ W <5 m), W_5_6 (5 m $\leq$ W <6 m), W_6_up (6 m $\leq$ W);
Length (L)	Chance node	L_below_7 (0 m $\leq$ L <7 m), L_7_11 (7 m $\leq$ L <11 m), L_11_up (11 m $\leq$ L);
Region Risk (R)	Chance node	R_low (region with low risk level), R_medial (region with medial risk level), R_high (region with high risk level);
Stall on LX (Stall)	Chance node	True, False;
Zigzag Violation (ZV)	Chance node	True, False;
PriC nodes		
Motorist Behavior Accident (MB)	Chance node	True, False;
Static Factor Accident (SF)	Chance node	True, False;
Consequence nodes		
SAL2 MV Accident (SA)	Chance node	True, False;
Fatalities (F)	Chance node	F_0 ($F=0$), F_0_up ($0<$ F);
Severe Injuries (S)	Chance node	S_0_2 ($0 \leq$ S <2), S_2_up ($2 \leq$ S);
Minor Injuries (M)	Chance node	M_0_3 ($0 \leq$ M <3), M_3_up ($3 \leq$ M);
Consequence Severity (CS)	Deterministic node	Level_1, Level_2, Level_3, Level_4, Level_5;

Table 6.5. CSCs for the BN risk model

ECs	(SA, F)$_e$, (SA, S)$_e$, (SA, M)$_e$, (SF, SA)$_e$, (MB, SA)$_e$, (CM, SF)$_e$, (RLS, SF)$_e$, (A, SF)$_e$, (P, SF)$_e$, (W, SF)$_e$, (L, SF)$_e$, (R, SF)$_e$, (ADRT, CM)$_e$, (ADRV, CM)$_e$, (Stall, MB)$_e$, (ZV, MB)$_e$, (B, Stall)$_e$, (Stop, Stall)$_e$;
FCs	(F, S)$_{bf}$, (F, M)$_{bf}$, (S, M)$_{bf}$, (W, L)$_{bf}$, (RSL, P)$_{bf}$, (RSL, A)$_{bf}$, (RSL, W)$_{bf}$, (RSL, L)$_{bf}$, (RSL, CM)$_{bf}$, (CM, A)$_{bf}$, (CM, P)$_{bf}$, (L, P)$_{bf}$, (L, A)$_{bf}$, (W, P)$_{bf}$, (W, A)$_{bf}$, (ADRV, ADRT)$_{bf}$, (Stop, B)$_{bf}$, (ADRT, B)$_{bf}$, (ADRT, Stop)$_{bf}$,(ADRT, Stall)$_{bf}$, (ZV, Stall)$_{bf}$, (ZV, B)$_{bf}$, (ZV, Stop)$_{bf}$, Between PriC nodes and TC nodes (bf), Between <F, S, M> and PriC nodes (bf), Between Consequence nodes and SC nodes (bf), Between Consequence nodes and TC nodes (bf);
PDCs	(CM, ZV)$_p$, (R, ZV)$_p$, (W, ZV)$_p$, (CM, B)$_p$, (A, B)$_p$, (L, Stop)$_p$, (W, Stop)$_p$;
bf: bidirectional forbidden;	

One can notice that, as shown in Fig. 6.8, the BN risk model contains two layers: 1) **Layer 1** is used for diagnosing influential factors; 2) **Layer 2** is used for evaluating consequences related to LX accidents. The "SAL2 MV Accident" node colored in yellow is the key node connecting the two layers, as well as the target node of accident prediction. In **Layer 1**, we split the network into 2 sub-networks: the static factor related network (SFN) and the motorist behavior factor related network (MBFN).

6.4.4 Model performance validation

Now that the BN structure is set up, we need to deal with model validation. In this section, ROC and AUC are adopted to evaluate the prediction performance of the present BN risk model. Regarding the AUC test, we should recall the following:

1) If AUC = 1, it is a perfect prediction model. When using it, a perfect prediction can be obtained with at least one threshold value.

2) If $0.5 < \text{AUC} < 1$, it is better than random guessing and has relatively sound predictive value.

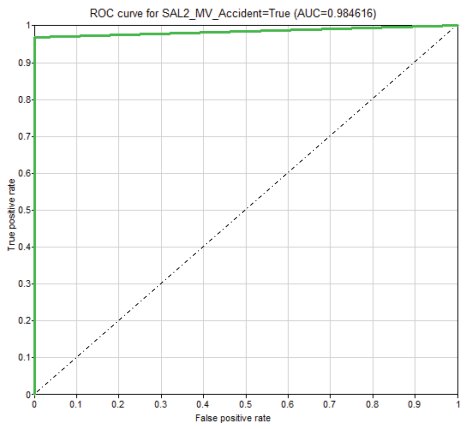
3) If AUC = 0.5, it is the same as random guessing, for example, throwing coin, thus, this model has no predictive value.

4) Otherwise, AUC < 0.5, it is worse than random guessing and valueless; but obviously, for the reverse-prediction, it is better than random guessing.

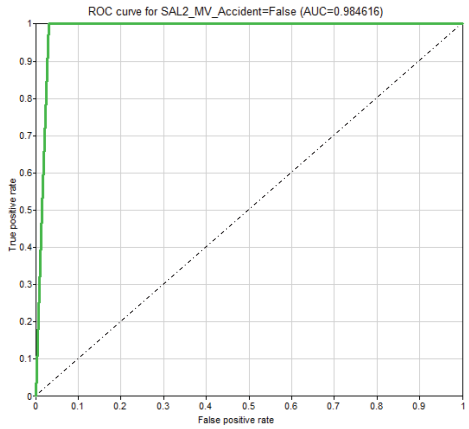
Therefore, one can notice that the ideal perfect ROC curve (cf. section 6.3) is the point (0, 1). Moreover, the closer the AUC to 1, the better the performance of a prediction model.

Besides, the K-fold cross-validation method is used to perform validation [GeNIe 1999] (the Deterministic node should be excluded when performing validation). Here, we set K=2, namely, the data set is divided into two parts of equal size (16664 samples) and the first part is used for parameters training, while the second part is used for validation. In our BN risk model, "SAL2 MV Accident", "Fatalities", "Severe Injuries" and "Minor Injuries" are the targeted prediction nodes which we care about. Therefore, we offer the ROC analysis of these 4 nodes in Fig. 6.5. The dash diagonal line indicates a baseline ROC curve for a predictor that is valuable when its real ROC curve is above this line. One can notice that the ROC curves respectively for "SAL2 MV Accident = True", "SAL2 MV Accident = False", "Minor Injuries = M_0_3", "Minor Injuries = M_3_up", "Severe Injuries = S_0_2", "Severe Injuries = S_2_up", "Fatalities = F_0" and "Fatalities = F_0_up" are all above the baseline. Moreover, the AUC values for the 8 states of the 4

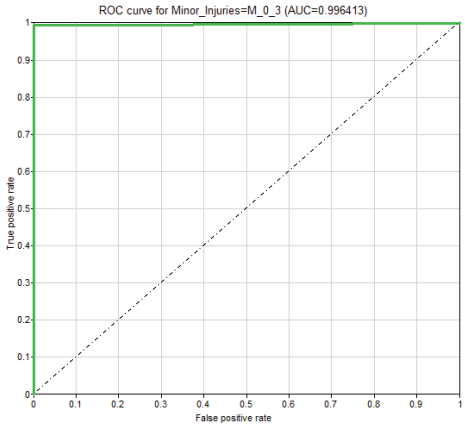
nodes are all between 0.9 and 1 (>> 0.5). Although the above results show significantly good performance for prediction of the BN model, we have been wondering whether this is not due to the fact that there are many "zero" accidents in reality. Therefore, a further analysis aimed to validate the model performance for prediction of "non-zero" accidents and their related consequence needs to be performed (refer to Table 6.7).



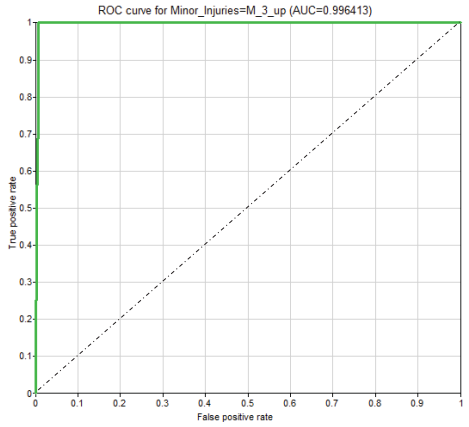
(a) The ROC for "SAL2 MV Accident = True"



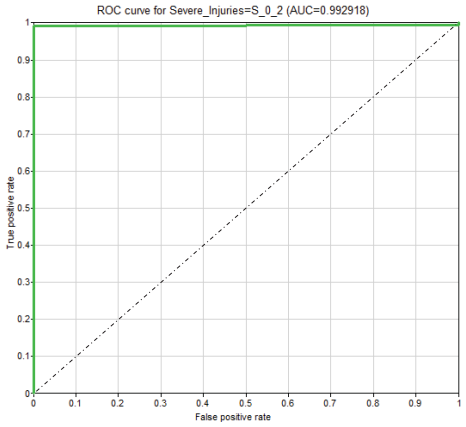
(b) The ROC for "SAL2 MV Accident = False"



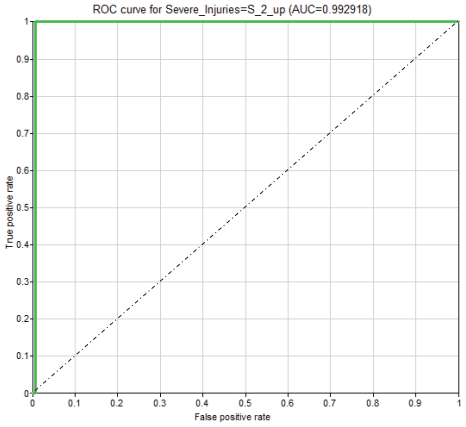
(c) The ROC for "Minor Injuries = M_0_3"



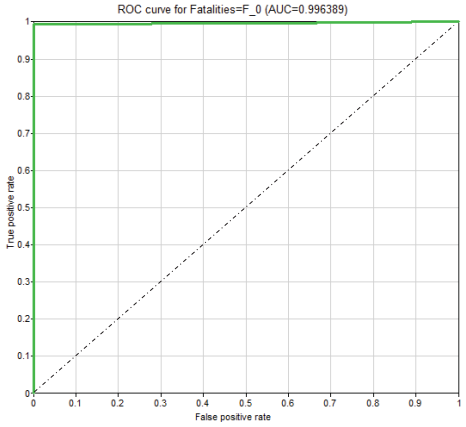
(d) The ROC for "Minor Injuries = M_3_up"



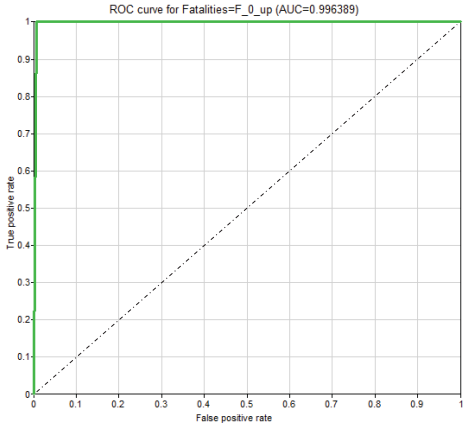
(e) The ROC for "Severe Injuries = S_0_2"



(f) The ROC for "Severe Injuries = S_2_up"



(g) The ROC for "Fatalities = F_0"



(h) The ROC for "Fatalities = F_0_up"

Fig. 6.5. The ROC curves for the 4 targeted nodes

Further comparison related to the prediction performance of the 4 nodes is performed between our BN model and the BN models automatically generated by BS, EGS, GTT, Naïve Bayes, ANB and TAN. As shown in Table 6.6, through investigating the results of the other 6 learning approaches, the entire accuracy and AUC values of our proposed model are clearly better than those of the other 6 learning approaches.

Table 6.6. Comparison of entire prediction performance

Approach	SA ACCU, AUC_T, AUC_F	F ACCU, AUC_0, AUC_{0_up}	S ACCU, AUC_{0_2}, AUC_{2_up}	M ACCU, AUC_{0_3}, AUC_{3_up}
Our model	0.9963, 0.9846, 0.9846	0.9801, 0.9964, 0.9964	0.9982, 0.9929, 0.9929	0.9913, 0.9963, 0.9963
BS	0.8751, 0.9187, 0.9187	0.8101, 0.5708, 0.5708	0.8638, 0.9541, 0.9541	0.8657, 0.9683, 0.9683
EGS	0.9134, 0.8857, 0.8857	0.9203, 0.8306, 0.8306	0.8509, 0.7790, 0.8509	0.8917, 0.8157, 0.8157
GTT	0.8706, 0.8216, 0.8216	0.8610, 0.8213, 0.8213	0.8704, 0.7126, 0.7126	0.8713, 0.8315, 0.8315
NB	0.6356, 0.5163, 0.5163	0.7704, 0.5856, 0.5856	0.8333, 0.6012, 0.6012	0.6181, 0.2015, 0.2015
ANB	0.9287, 0.9015, 0.9015	0.9516, 0.9340, 0.9340	0.9287, 0.9111, 0.9111	0.9414, 0.9202, 0.9202
TAN	0.9539, 0.9431, 0.9431	0.9636, 0.9616, 0.9616	0.9891, 0.9680, 0.9680	0.9847, 0.9794, 0.9794

ACCU: accuracy;

Moreover, the prediction accuracy for accident/consequence occurrence is investigated to further compare the prediction performance between our model and the 6 traditional learning approaches. As shown in Table 6.7, the accuracy values for "SA = False"/"SA = True" (1/0.9622), "F = 0"/"F = 0_*up*" (1/0.9020), "S = 0_2"/"S = 2_*up*" (1/0.6) and "M = 0_3"/"M = 3_*up*" (1/0.75) of our model are relatively higher than those of the other 6 learning approaches. Note that the sample size of single accident related to "severe injuries more than 2" and "minor injuries more than 3" is small in reality, which lead to the lower accuracy compared with the accuracy of "SA = True" and "F = 0_*up*".

Table 6.7. Comparison of prediction performance for accident/consequence occurrence

Approach	SA		F		S		M	
	$ACCU_{False}$, $ACCU_{True}$		$ACCU_0$, $ACCU_{0\_up}$		$ACCU_{0\_2}$, $ACCU_{2\_up}$		$ACCU_{0\_3}$, $ACCU_{3\_up}$	
Our model	1	0.9622	1	0.9020	1	0.6	1	0.75
BS	0.8815	0.7181	0.9101	0.6728	0.9168	0	0.9674	0.1250
EGS	0.9242	0.6875	0.9233	0.7326	0.9615	0	0.9510	0.1250
GTT	0.9562	0.7162	0.9410	0.7813	0.9704	0	0.9613	0.1250
NB	0.6905	0.5637	0.7514	0.5576	0.7334	0	0.7182	0
ANB	1	0.8011	0.9616	0.8341	0.9617	0.2000	0.9202	0.2500
TAN	0.9693	0.8437	0.9561	0.6614	0.9238	0	0.9742	0

ACCU: accuracy;

The better performance of our proposed model is mainly attributed to the incorporation of expert knowledge and preliminary causality identification. Indeed, this significantly reduces the negative effect of trivial correlations and improves the reliability of the identified causal relationships among the variables considered. Therefore, these validation results indicate that our BN risk model has relatively sound prediction performance and allow us to consider the outcomes of the model to be trustworthy. Besides, this attests that the proposed BNI-RR framework promotes the efficiency of risk analysis.

6.5 Analysis and discussion

In this section, we will illustrate how the BNI-RR framework can be advantageously worked out to perform risk analysis on LXs. We should mention that the aspects discussed in the sequel do not represent the exhaustive capabilities through our framework, and should be regarded as illustrations.

6.5.1 Forward and reverse inferences

Based on the BN risk model, one can estimate the probability of a train-MV accident occurring at an SAL2 LX through forward inference. As shown in Fig. 6.6, the general probability of a train-MV accident occurring at an SAL2 over the four years influenced by the interaction of all the factors considered, is estimated as almost 0.0061. As an illustration, the probability of a train-MV accident caused by static factors is about 0.0011 and the probability of a train-MV accident caused by inappropriate motorist behavior is about 0.0049. Moreover, fatalities and severe injuries caused by the accident are, to a large extent, fewer than 1 and 2, respectively ($P(F = F\_0) = 0.9993$, $P(S = S\_0\_2) = 0.9999$). Minor injuries caused by an SAL2 accident are most likely to be fewer than 3 ($P(M = M\_0\_3) = 0.9998$). Thus, the consequence severity level are most likely to be level 1 ($P(CS = Level\_1) = 0.9990$).

Fig. 6.7 shows that the probability of a train-MV accident occurring at an SAL2 would increase to 0.0384 if all the risky states of secondary causes occur, namely "Corrected Moment" in the "CM_49_up" group, "Railway Speed Limit" in the "RSL_110_up" group, "Alignment" in the "S_shape" group, "Profile" in the "Hump_cavity" group, "Width" in the "W_6_up" group, "Length" in the "L_11_up" group, "Region Risk" in the "R_high" group, "Stall on LX" being true and "Zigzag Violation" being true. The related consequences are likely to be severer as well. In this way, various prediction results for the targeted nodes in terms of various combinations of the different states of the other impacting factors can be obtained through forward inference. Here, we do not exhaust all

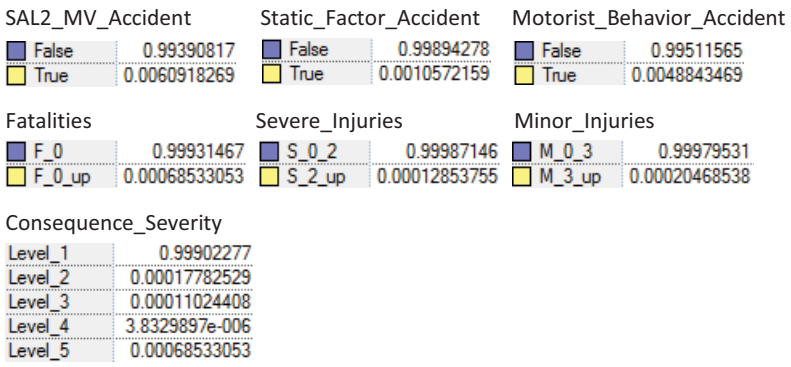


Fig. 6.6. General prediction

prediction results due to limited space.

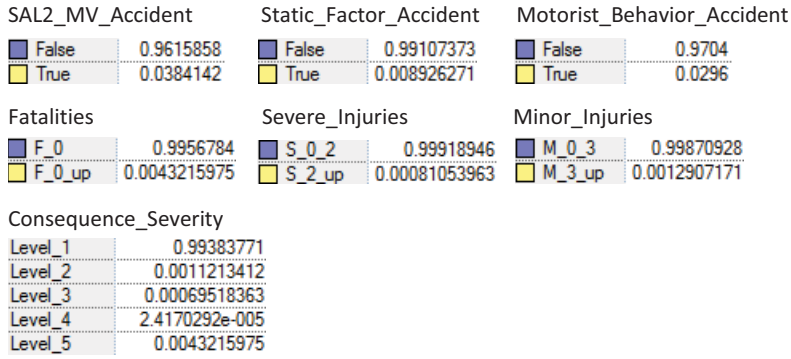


Fig. 6.7. Prediction related to the occurrence of severest states of secondary causes

Subsequently, the "SAL2 MV Accident = True" state is configured as the targeted state. In this way, one can assess the contribution degree of each influential factor to train-MV accident occurrence through reverse inference. Detailed results are given in Fig. 6.8. It is worth noticing that accidents caused by inappropriate motorist behavior contribute 80% to the entire train-MV accidents at SAL2 LXs, while accidents caused by static factors contribute only 17%. As for inappropriate motorist behavior, "Zigzag Violation" is more significant than "Stall on LX" in terms of causing train-MV accidents, because of the contribution of 58% (compared with 42% contribution of "Stall on LX"). On the other hand, in terms of static factors, when a train-MV accident occurs at an SAL2 LX, this LX has the probabilities of 74%, 38%, 44%, 37% and 46% respectively involved in the most risky situations that "Corrected Moment" in the "CM_49_up" group, "Railway Speed Limit" in the "RSL_110_up" group, "Width" in the "W_6_up" group, "Length" in the "L_11_up" group and "Region Risk" in the "R_high" group. These results indicate that more attention needs to be paid to LXs having the above risky characteristics.

Moreover, special accommodation and/or technical solutions need to be implemented to prevent motorist zigzag violations, for instance, transforming SAL2 LXs into SAL4 LXs (Four-half barrier systems), or installing median separators between opposing lanes of road traffic in front of SAL2 LXs. As for the consequences caused by accident, it is most likely to be 0 fatality ($P(F = F\_0) = 0.8875$), less than 2 severe injuries ($P(S = S\_0\_2) = 0.9789$) and less than 3 minor injuries($P(M = M\_0\_3) = 0.9664$). Thus, to a large extent, the consequence severity would be Level 1 ($P(CS = Level\_1) = 0.8396$, $P(CS = Level\_2) = 0.0292$, $P(CS = Level\_3) = 0.0181$, $P(CS = Level\_4) = 0.0006$ and $P(CS = Level\_5) < 0.1125$). Hence, one can set various states of the consequence nodes as the targeted states to make thorough corresponding diagnosis of causal factors through reverse inference.

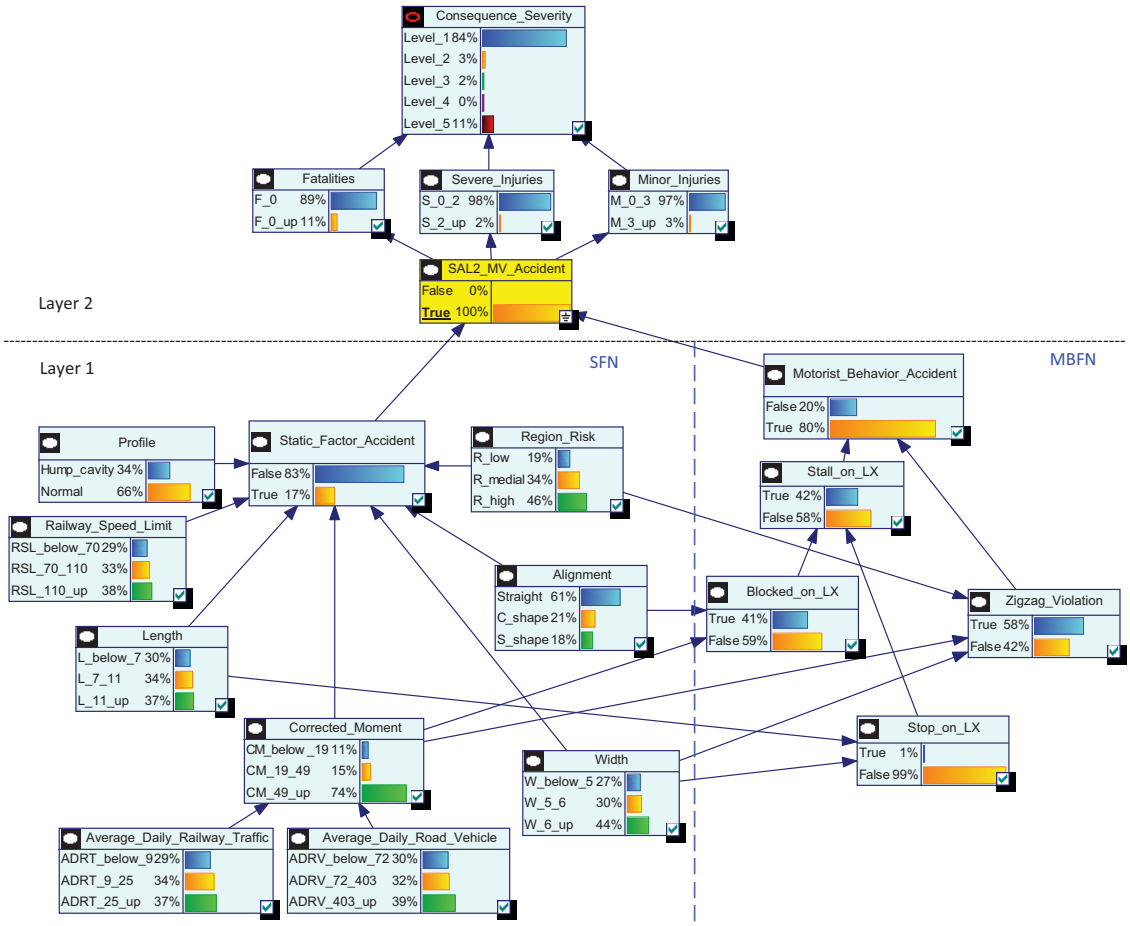


Fig. 6.8. Cause diagnosis when a train-MV accident occurs

6.5.2 Influence and sensitivity analysis

Based on the BN risk model, the influence strength (IS), which represents the impact level of parent nodes on their respective child nodes, can be computed through GeNIe tool using the Euclidean distance [Koiter 2006] which is commonly used to measure the influence strength between parent and child nodes. Fig. 6.9 shows the normalized influence strength (labeled on the arcs) between each pair of parent and child nodes. Fatalities (1) has stronger influence on consequence severity than severe injuries (0.5000) and minor injuries (0.4800). Moreover, inappropriate motorist behavior (0.7099) impacts more on LX accident occurrence than static factors (0.2899). Among the static factors, the region risk factor (0.0010) has the greatest impact on LX accident occurrence, while the profile (0.0002) has the slightest impact on LX accident occurrence. On the other hand, zigzag violation (0.1143) impacts more on LX accident occurrence caused by inappropriate motorist behavior, compared with the impact of "stall on LX" (0.0137).

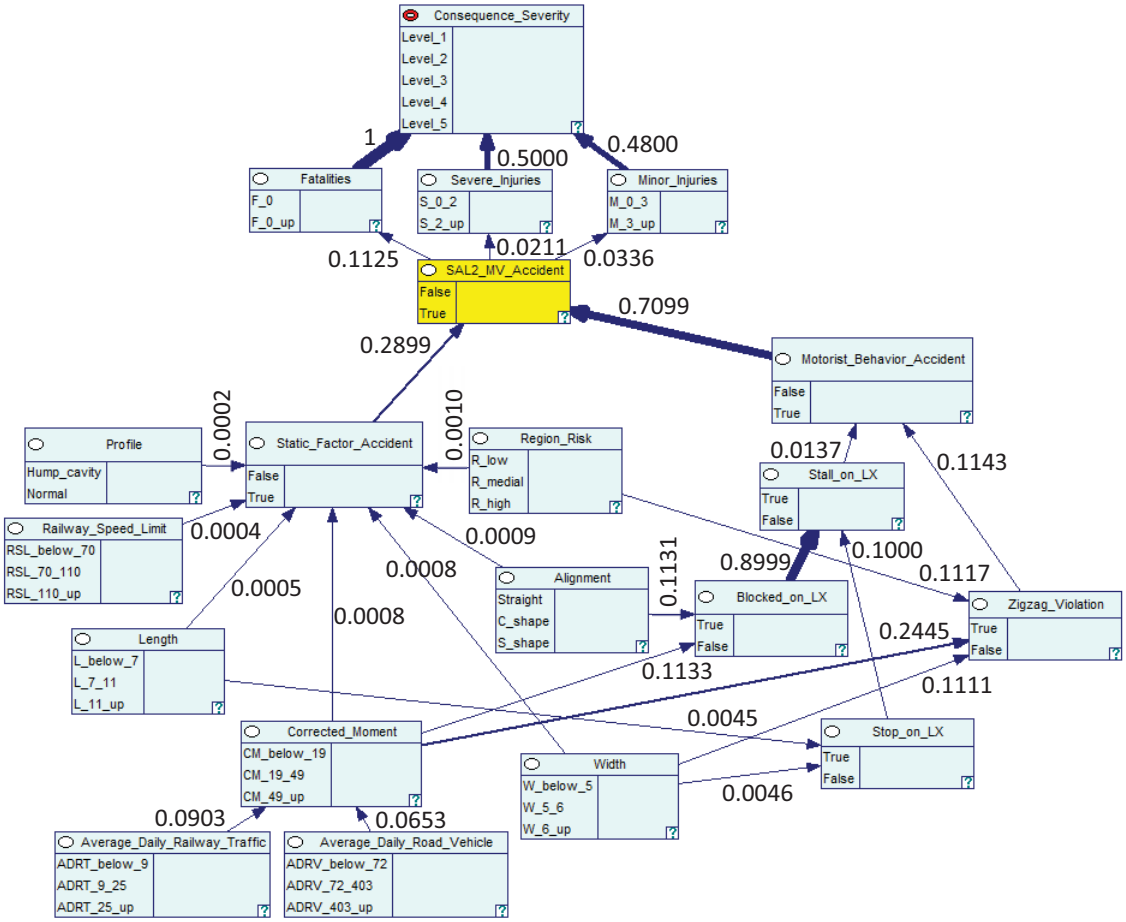


Fig. 6.9. Normalized influence strength analysis

Furthermore, the sensitivity analysis is performed to interpret the sensitivity of $P(SA = True)$ to various conditional probabilities of different variable combinations. The Sensitivity Tornado Diagram (STD) is shown in Fig. 6.10. The top horizontal axis represents the values of $P(SA = True)$. The vertical axis represents the general prediction value of $P(SA = True) \approx 0.0061$, which is set as the vertical datum axis. The horizontal bars are viewed as two parts divided by the vertical datum axis. Green bars represent the values of $P(SA = True)$ decreasing from the datum value while red bars represent the values of $P(SA = True)$ increasing from the datum value, according to the changes of impacting conditional probabilities P_c. As the values of P_c change within the interval $[P_c - 0.1P_c, P_c + 0.1P_c]$ (setting the spread degree as 0.1), the values of $P(SA = True)$ change within an interval $[P(SA = True)_{min}, P(SA = True)_{max}]$ and distribute with respect to the changing values of P_c in the whole range $[P_c - 0.1P_c, P_c + 0.1P_c]$ accordingly. Here, the values of $P(SA = True)$ change within the interval $[0.0045, 0.1055]$.

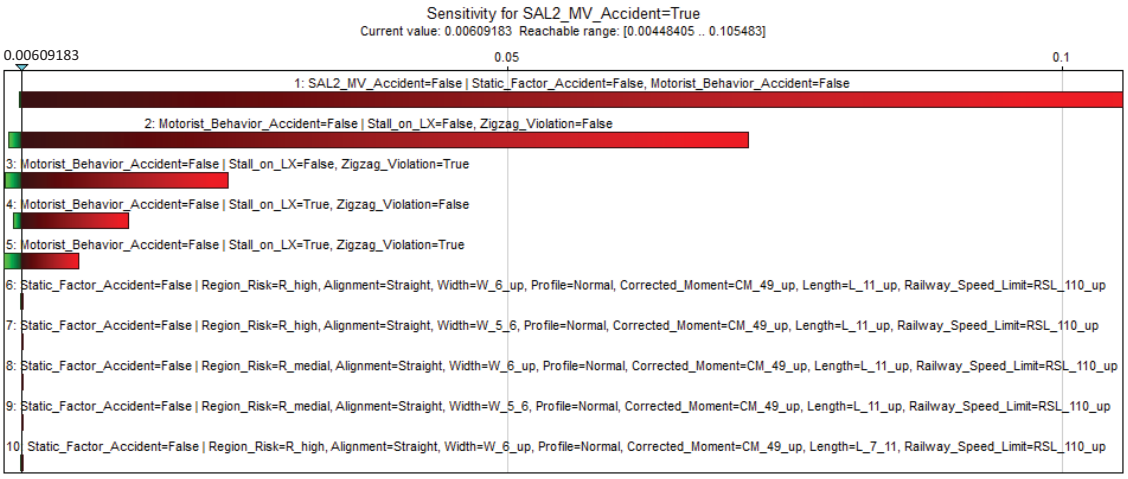


Fig. 6.10. Sensitivity tornado diagram

Fig. 6.10 shows the top 10 impacting P_cs which $P(SA = True)$ is most sensitive to. One can notice that $P(SA = False|SF = False, MB = False)$ impacts $P(SA = True)$ most. Namely, $P(SA = True)$ decreases from 0.1055 to 0.0059 as $P(SA = False|SF = False, MB = False)$ increases from 0.8999 to 1. As for the P_cs related to motorist behavior, $P(MB = False|Stall = False, ZV = False)$ (taking the second place) impacts $P(SA = True)$ most, compared with the other P_cs related to motorist behavior. As for the P_cs related to static factors, $P(SA = True)$ is most sensitive to $P(SF = False|R = R\_high, A = Straight, W = W\_6\_up, P = Normal, CM = CM\_49\_up, L = L\_11\_up, RSL = RSL\_110\_up)$ (taking the sixth place), compared with the other P_cs related to static factors. These results further attest that the LX acci-

dent occurrence is more sensitive to inappropriate motorist behavior than static factors. Moreover, as for motorist behavior factors, the LX accident occurrence is most sensitive to zigzag violation occurrence, compared with other motorist behavior factors. On the other hand, as for static factors, the LX accident occurrence is most sensitive to the riskiest states of various static factors. Therefore, the improvement measures need to be targeted on mitigating the above high-sensitivity factors, since a small scale of improvement in such factors can potentially reduce the LX risk as a whole on a large scale.

6.6 Summary

In this chapter, an effective and comprehensive modeling framework for risk reasoning, called BNI-RR, is proposed, which consists of a set of integrated processes, namely, risk scenario definition, real field data collection and processing, BN model establishment and model performance validation.

Then, the BNI-RR framework is applied to LX risk analysis. The output of our study offers a valuable support for decision making regarding LX safety. Although the BNI-RR framework is applied to the risk analysis of French LXs in our study, this framework is a general approach that can be applied to different contexts related to risk analysis.

According to the aforementioned analyses, the main contributions of the present study are as follows:

1) A causal semantics definition is proposed to describe the DAG of BN, which consists of three elements, namely, $IF, THEN$ and CAK. Thus, causal structural constraints are introduced based on the concept of CAK for the purpose of causality optimizing. With the help of causal structural constraints, expert knowledge can be integrated to distinguish causalities from correlations. Therefore, inappropriate connections are neglected so as to facilitate highlighting the main causes leading to LX accidents.

2) Based on the causal BN model, we were able to make forward inference and reverse inference, which are two valuable complementary means for performing inductive and deductive diagnosis. For instance, our BN risk model allows us not only to predict the probability of accident occurrence, but also evaluate the related consequence severity level, quantify the respective contribution degrees of various factors to the overall risk and identify the riskiest factors. These aspects are rarely achieved in existing related works and demonstrate the effectiveness of utilizing our BNI-RR framework.

3) Influence strength analysis and sensitivity analysis are two further approaches that were adopted to finely investigate the influence strength of causal factors on conse-

quence factors and determine which causal factors the consequence factors are most sensitive to. Based on the obtained results, adequate targeted technical/societal solutions/decisions and improvement recommendations can be identified to act on specific causal factors.

To sum up, the aforementioned contributions show that the BNI-RR approach offers an integrated modeling and analysis framework that allows for performing thorough risk analyses on a given LX or a set of LXs at a global level. The findings obtained through applying the BNI-RR framework on LX risk analysis offer a significant perspective on the major factors causing LX accidents and pave the way for identifying practical design measures and improvement recommendations to prevent accidents at LXs.

Part III

OVERALL CONCLUSIONS AND PERSPECTIVES

Chapter 7

Overall Conclusions and Perspectives

Sommaire

7.1 Overall Conclusions

This dissertation focuses on the analysis of LX risk through various advanced techniques. Our contributions are discussed in Part II which consists of chapter 3 dedicated to preliminary statistical analysis on LXs, chapter 4 that focuses on motorist behavior quantitative analysis, chapter 5 that investigates advanced statistical accident prediction modeling, chapter 6 dedicated to Bayesian network based framework for LX risk reasoning.

In chapter 3, a general risk analysis of average accident frequency in terms of transport mode and geographical region is performed. Then, the risk analysis in terms of frequency coefficient, namely the average accident frequency acted by the traffic moment, is performed with regard to various traffic moment groups. Finally, the frequency coefficient distributed in different French regions is investigated. To sum up, we have assessed the effect of the above factors on the risk level quantitatively, in such a way as to open the way for setting efficient solutions to improve LX safety.

In chapter 4, a risk analysis of motorist behavior is performed based on field measurement conducted at 12 automatic LXs (11 equipped with two half barriers (SAL2) and 1 equipped with four half barriers (SAL4)) while distinguishing between different phases of LX closure cycle. The global violation trend of motorist behavior during Ph2 and Ph3 of LX closure cycle is investigated firstly. Then, an analysis on the violation rate during Ph2 and Ph3 according to the week and the hour is performed. Furthermore, an analysis on the speed of violating vehicles during Ph2 and Ph3 is performed. As for the motorist behavior during Ph4, we analyze the impact of prolonged Ph4 duration on the zigzag violation rate of motorists. Moreover, the impact of LX location (in terms of proximity to a railway station) on the zigzag violation rate of motorists during Ph4 is analyzed.

Subsequently, the impact of road traffic density on the waiting queue in front of LXs and troop phenomenon is investigated. Finally, the comparison of motorist responses to SAL2 and SAL4 LXs is performed.

In chapter 5, an advanced accident frequency prediction model, which enables to rank risky LXs accurately and identify the significant impacting parameters efficiently, is developed. In this model, we take into account the corrected traffic moment which is more effective in estimating automobile-involved LX accidents frequency compared with the conventional traffic moment, single average daily railway traffic or single average daily road traffic. The impact of road accident factor is highlighted in this model, which was likely to be ignored in previous studies related to LX safety analysis. Moreover, we originally introduce the region LX-accident-prone factor in our study to interpret the variation of LX accident statistics with regard to various regions. One can notice that the risk of LX accidents is most sensitive to the region LX-accident-prone factor, compared with other LX characteristics. In fact, the Nonlinear Least-Squares (NLS) method, Poisson regression method, negative binomial (NB) regression method, zero-inflated Poisson (ZIP) regression method and zero-inflated negative binomial (ZINB) regression method have been employed to estimate the respective coefficients of parameters in the prediction model. Then, a validation process is performed based on various statistical and probabilistic means to examine how well the estimation of the model fits the reality. The validation results attest that the NB distribution combined with λ_{10Y} shows relatively higher prediction performance than other combinations. Finally, a comparison between our present model and two existing reference models is carried out and allows showing the good efficiency of our model. Since this prediction model can estimate the LX accident frequency which reflects the LX risk level, it is possible to rank LXs according to their risk level and, thereby, pay more attention on the riskiest LXs. Moreover, with this model, we can predict the probability of the exact number of accidents occurring at a given LX during a long time period, thus to estimate the corresponding losses and investment in countermeasures. We believe that this represents a valuable decision-making support to road and railway stakeholders.

In chapter 6, a BN based framework for causal reasoning related to risk analysis is proposed. It consists of a set of integrated stages, namely risk scenario definition, real field data collection and processing, BN model establishment and model performance validation. Causal structural constraints are introduced to the framework for the purpose of combining expert knowledge with data-driven, thus to identify effective causalities and avoid inappropriate structural connections. The BN risk model is established based on the real field data of LX accidents/incidents and the model performance is validated. The validation results indicate that our model has high accuracy in terms of LX risk prediction. Then, forward inference and reverse inference based on the BN risk model are performed to

predict LX accident occurrence and quantify the contribution degree of various impacting factors respectively, so as to identify the riskiest factors. Finally, influence strength and sensitivity analyses are further carried out to scrutinize the influence strength of various causal factors on the LX accident occurrence and determine which factors the LX accident occurrence likelihood is most sensitive to. With this BN model, one can assess the gain brought by the improvement of one or more factors considered at a given LX in terms of accident occurrence and corresponding consequences, as well as the further investment. In this respect, this model can be regarded as a support for decision-making.

In summary, the study reported in this dissertation offers an in-depth perspective on LX risk analysis, as explained through the above contributions. Especially for practicing risk managers and decision makers, our study provides a thorough quantitative analyses pertaining to LX risk and paves the way for identifying practical design measures and technical recommendations to improve LX safety.

7.2 Perspectives

One should notice that the work presented in this dissertation still shows some limitations. For example, due to the lack of available detailed data, the present study does not consider all the potential impacting factors pertaining to LX safety, for example, visibility, weather, etc. An additional limitation of the current study is the intrinsic unpredictability of motorist behavior. Even though certain outcomes are correlated with certain human characteristics, motorist behavior is a moving target and it may change over time accompanying the changes in terms of social lifestyle, regulation or policies. For example, the operation of vehicles fitted with new facilities, such as autonomous cars, would also have a strong impact on risk assessment in the future. Hence, some updates are continuously needed to take into account such potential variations. By the way, this limitation is a quite general one rather than a specific feature of our experiments, since it basically applies to any experimental study related to human behavior as well.

The work discussed in this dissertation raises several research directions as discussed below:

1. Regarding further impacting parameters

   In future works, we will investigate further impacting factors (e.g., region road accident factor) to better understand the aspects which stimulate motorist violations at LXs and bring into play Bayesian risk models to quantify the causal relationships between the factors and accident occurrence, so as to assess their impacts on the whole risk level. Our risk model will be enriched while integrating further factors

according to real data and causality analysis that we will perform on the basis of fault events. In addition, the analysis made on the basis of Bayesian network can be extended by means of Valuation-Based Systems to deal with uncertain parameters and/or imprecise probabilities.

2. Regarding dynamic BN models

   Dynamic BNs [Murphy 2002] will be considered for further modeling the dynamic/temporal aspects of some of the considered constraints nodes (railway traffic, road traffic, etc.) to improve the risk analysis prediction, while the time-dependent conditional probabilities can be obtained. However, this technique shall be performed carefully, as the changes introduced by its application could increase exponentially the complexity of the model and the necessary calculation time. In addition, since inappropriate motorist behavior has been identified as the main cause of LX accidents according to our BN model, a thorough analysis of this issue combining both qualitative and quantitative techniques should be carried out to determine the adequate countermeasures.

3. Regarding further investigation on the reasons for inappropriate motorist behavior

   As mentioned in 2), inappropriate motorist behavior is the main factor for LX accidents. In fact, the causes of LX accidents related to motorist behavior are complex and we could divide them into three main mechanisms:

   (1) The motorist driving error: the motorist fails to analyze the road situation correctly, in the hierarchy of dangers. For instance, he/she is afraid that his/her vehicle will rub on the decking or touch the overhead grid and he/she will stop in the danger LX intersection zone or he/she judges inappropriately the length or speed of a waiting queue in front of LX. This mechanism seems to cause about 30% [SNCF 2017] of LX accidents. It results in few road deaths, because the immobilization of the vehicle can take place before the arrival of the train, often even before closing the barriers, and the driver may have left the vehicle before the collision occurs. However, these accidents can be very serious in terms of railway consequences, especially, if in case of heavy vehicle (trucks, trailers, etc.).

   (2) The “looked but not see”, in other words: the motorist does not see or notice the LX until it is too late. He can brake suddenly, then stall on the LX or slip and end up with a wheel of his vehicle in the ballast, or even hitting the barriers and the train.

   This mechanism is also responsible for about 30% [SNCF 2017] of LX accidents. Moreover, collisions occurring during the closure cycle of the LX are often deadly,

since the motorists and passengers are not able to escape from the vehicle. Therefore, it is important to make LXs more conspicuous, thus to avoid "looked but not see" accidents.

In addition, making LX environment simpler and avoiding left turns to prevent non-deliberate violations is still significant.

(3) The intentional zigzag: zigzag crossings often involve light vehicles, two-wheel vehicles or pedestrians. These zigzags are the main reason for a number of road deaths as we analyzed in chapter 4.

We wish to investigate the detailed reasons behind motorist behavior using an integrated methodology in the future. Namely, we believe that utilizing jointly vehicle dynamics recording, a systems approach and a psychological schema theory can provide a more thorough understanding of the reasons underlying motorist violations.

4. Regarding justification of investment in improving LX safety

   Based on our analyses, design measures and technical recommendations to improve LX safety can be identified. Further experiments need to be carried out to investigate the efficiency of some technical solutions (e.g., installing LED strobe or rumble strip, adding median separator, transforming SAL2 LX into SAL4 LX, adding obstacle detector, etc.).

   Moreover, from this point of view, the investment in these solutions needs to be assessed in terms of cost efficiency. In the future, Influence Diagrams (IDs) [Vatn 2002], which are extensions of BNs, can be used to assess such investment. An example of IDs is shown in Fig. 7.1. In this ID, additional nodes for decisions (rectangles) and utilities (diamond shape) are attached to the BNs. The directed links (arrows) represent probabilistic dependencies among the system variables (represented by X), decision variables (represented by a) and utility variables (represented by U) in the network. The decision nodes are introduced as parents to the chance nodes and the utility nodes. With given data, the investment can thus be evaluated through utility nodes correspondingly according to the change of configuration in decision nodes.

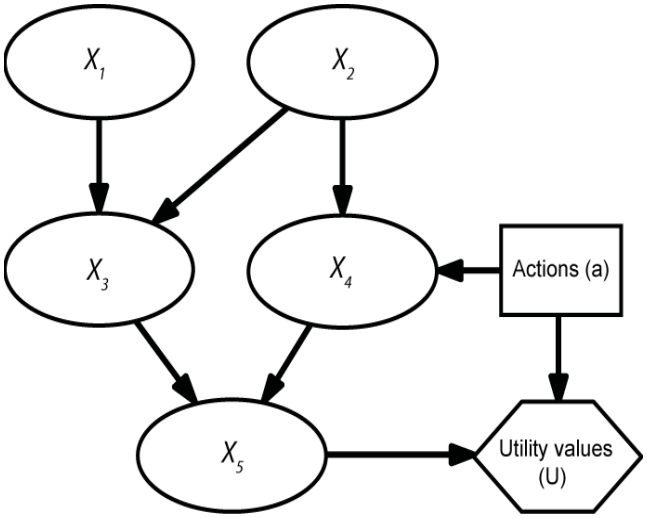


Fig. 7.1. An example of IDs

Bibliography

[50126 1999] EN 50126. Railway applications-the specification and demonstration of reliability, availability, maintainability and safety (rams). British Standards Institution, 1999. (Cited in page 146.)

[Abraham 1998] J. Abraham, T. Datta et S. Datta. *Driver behavior at rail-highway crossings.* Transportation Research Record: Journal of the Transportation Research Board, no. 1648, pages 28–34, 1998. (Cited in page 14.)

[Administration 2012] Federal Railroad Administration. Highway/rail incidents summary tables. 2012. (Cited in pages 4 and 87.)

[Anandarao 1998] S. Anandarao et C. D. Martland. *Level crossing safety on East Japan Railway Company: Application of probabilistic risk assessment techniques.* Transportation, vol. 25, no. 3, pages 265–286, 1998. (Cited in page 82.)

[Anscombe 1973] F. J. Anscombe. *Graphs in statistical analysis.* The American Statistician, vol. 27, no. 1, pages 17–21, 1973. (Cited in page 94.)

[ATC 2010] ATC. *National Railway Level Crossing Safety Strategy 2010-2020.* Main Roads Western Australia, Australia, 2010. (Cited in page 82.)

[Austin 2002] R. D. Austin et J. L. Carson. *An alternative accident prediction model for highway-rail interfaces.* Accident Analysis & Prevention, vol. 34, no. 1, pages 31–42, 2002. (Cited in pages 19, 90 and 119.)

[Bearfield 2005] G. Bearfield et W. Marsh. *Generalising event trees using Bayesian networks with a case study of train derailment.* Dans International Conference on Computer Safety, Reliability, and Security, pages 52–66. Springer, 2005. (Cited in page 129.)

[Bensi 2010] M.T. Bensi. A bayesian network methodology for infrastructure seismic risk assessment and decision support. University of California, Berkeley, 2010. (Cited in page 35.)

[Borkowf 2002] C. B. Borkowf. *Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation.* Computational statistics & data analysis, vol. 39, no. 3, pages 271–286, 2002. (Cited in page 90.)

[Bouillaut 2013] . Bouillaut, O. Francois et S. Dubois. *A Bayesian network to evaluate underground rails maintenance strategies in an automation context.* Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, vol. 227, no. 4, pages 411–424, 2013. (Cited in page 34.)

[Bozdogan 1987] H. Bozdogan. *Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions.* Psychometrika, vol. 52, no. 3, pages 345–370, 1987. (Cited in page 106.)

[Buddhavarapu 2016] P. Buddhavarapu, J. G. Scott et J. A. Prozzi. *Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data.* Transportation Research Part B: Methodological, vol. 91, pages 492–510, 2016. (Cited in pages 21 and 98.)

[Bureau 2012] Australian Transport Safety Bureau. *Australian Rail Safety Occurrence Data: 1 July 2002 to 30 June 2012 (ATSB Transport Safety Report RR-2012-010).* Canberra, Australia: ATSB, 2012. (Cited in page 4.)

[Carlson 1999] P. Carlson et K. Fitzpatrick. *Violations at gated highway-railroad grade crossings.* Transportation Research Record: Journal of the Transportation Research Board, no. 1692, pages 66–73, 1999. (Cited in page 14.)

[Castillo 1997] E. Castillo, J.M. Gutiérrez et A.S. Hadi. *Sensitivity analysis in discrete Bayesian networks.* IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 27, no. 4, pages 412–423, 1997. (Cited in page 138.)

[Chadwick 2014] S. G. Chadwick, N. Zhou et M. R. Saat. *Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the US.* Safety Science, vol. 68, pages 128–137, 2014. (Cited in pages 18, 82 and 87.)

[Chang 2005] L. Y. Chang. *Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network.* Safety science, vol. 43, no. 8, pages 541–557, 2005. (Cited in pages 21 and 98.)

[Chatterjee 2015] S. Chatterjee et A. S. Hadi. Regression analysis by example. John Wiley & Sons, 2015. (Cited in page 94.)

[Chemweno 2015] P. Chemweno, L. Pintelon, Van H.A. et P. Muchiri. *Development of a risk assessment selection methodology for asset maintenance decision making: An analytic network process (ANP) approach.* International Journal of Production Economics, vol. 170, pages 663–676, 2015. (Cited in page 34.)

[Cheng 1997] J. Cheng, D.A. Bell et W. Liu. *An algorithm for Bayesian belief network construction from data.* Dans proceedings of AI & STAT'97, pages 83–90, 1997. (Cited in pages 133 and 137.)

[Committee 2012] National ALCAM Committee. *ALCAM in Detail-An Introduction to the new ALCAM models.* Australia, 2012. (Cited in pages 20 and 88.)

[Cooper 1990] G.F. Cooper. *The computational complexity of probabilistic inference using Bayesian belief networks.* Artificial intelligence, vol. 42, no. 2-3, pages 393–405, 1990. (Cited in page 133.)

[Cooper 2007] D. L Cooper, K. E. MacLeod et D. R. Ragland. *Rail crossings: a strategy to select countermeasure improvements for rail-highway crossings in California.* Safe Transportation Research & Education Center, 2007. (Cited in pages 14 and 137.)

[Cox 1994] E. Cox. *The fuzzy systems handbook: a practitioner's guide to building, using, maintaining fuzzy systems. Boston: AP Professional.* Rapport technique, ISBN 0-12-194270-8, 1994. (Cited in page 32.)

[Dagum 1993] P. Dagum et M. Luby. *Approximating probabilistic inference in Bayesian belief networks is NP-hard.* Artificial intelligence, vol. 60, no. 1, pages 141–153, 1993. (Cited in page 133.)

[Dai 2013] H. Dai, Y. Bao et M. Bao. *Maximum likelihood estimate for the dispersion parameter of the negative binomial distribution.* Statistics & Probability Letters, vol. 83, no. 1, pages 21–27, 2013. (Cited in page 101.)

[Dash 1999] D. Dash et M.J. Druzdzel. *A hybrid anytime algorithm for the construction of causal models from sparse data.* Dans Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 142–149. Morgan Kaufmann Publishers Inc., 1999. (Cited in page 137.)

[Davey 2005] J. Davey, N. Ibrahim et A. Wallace. *Motorist behaviour at railway level crossings: An exploratory study of train driver experience.* In Proceedings of the Road Safety Research, Policing and Education Conference, 2005. (Cited in page 82.)

[Davey 2008] J. Davey, A. Wallace, N. Stenson et J. Freeman. *The experiences and perceptions of heavy vehicle drivers and train drivers of dangers at railway level crossings.* Accident Analysis & Prevention, vol. 40, no. 3, pages 1217–1222, 2008. (Cited in pages 15, 16 and 17.)

[Dawid 1992] A.P. Dawid. *Applications of a general propagation algorithm for probabilistic expert systems.* Statistics and computing, vol. 2, no. 1, pages 25–36, 1992. (Cited in page 133.)

[De Campos 2007] L.M. De Campos et J.G. Castellano. *Bayesian network learning algorithms using structural restrictions.* International Journal of Approximate Reasoning, vol. 45, no. 2, pages 233–254, 2007. (Cited in page 142.)

[Dempster 1977] A.P. Dempster, N.M. Laird et D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977. (Cited in page 136.)

[ERA 2014] ERA. *Railway safety performance in the European Union.* 9 (2) Agency Regulation 881/2004/EC, 2014. (Cited in pages 4 and 40.)

[ERA 2016] ERA. *Railway safety performance in the European Union.* Rapport technique, doi:10.2821/129870, 2016. (Cited in page 13.)

[Ericson 1999] C.A. Ericson. *Fault tree analysis*. Dans System Safety Conference, Orlando, Florida, pages 1–9, 1999. (Cited in pages 23, 24 and 129.)

[Ericson 2015] C.A. Ericson*et al.* Hazard analysis techniques for system safety. John Wiley & Sons, 2015. (Cited in page 27.)

[E.U. 2012] E.U. *EU transport in figures: statistical pocketbook*. Dans Luxembourg: European Union, 2012. (Cited in page 13.)

[FHWA 1991] FHWA. *Warning time requirements at railroad-highway grade crossings with active traffic control*. Report No. FHWA SA-91-007, Federal Highway Administration. Washington, DC: February, 1991. (Cited in page 80.)

[Friedman 1997] N. Friedman, D. Geiger et M. Goldszmidt. *Bayesian network classifiers*. Machine learning, vol. 29, no. 2-3, pages 131–163, 1997. (Cited in page 137.)

[Fung 1994] R. Fung et B. Del Favero. *Backward simulation in Bayesian networks*. Dans Proceedings of the Tenth international conference on Uncertainty in artificial intelligence, pages 227–234. Morgan Kaufmann Publishers Inc., 1994. (Cited in page 133.)

[Gagniuc 2017] P.A. Gagniuc. Markov chains: From theory to implementation and experimentation. John Wiley & Sons, 2017. (Cited in page 29.)

[GeNIe 1999] GeNIe. Genie modeler user manual, version 2.1.1. From http://support.bayesfusion.com/docs/, 1999. (Cited in pages 131 and 150.)

[Ghazel 2009] M. Ghazel. *Using stochastic Petri nets for level-crossing collision risk assessment*. IEEE transactions on intelligent transportation systems, vol. 10, no. 4, pages 668–677, 2009. (Cited in page 4.)

[Ghazel 2014] M. Ghazel et E. M. El-Koursi. *Two-half-barrier level crossings versus four-half-barrier level crossings: A comparative risk analysis study*. IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 3, pages 1123–1133, 2014. (Cited in pages 4, 32, 81 and 129.)

[Ghazel 2017] M. Ghazel. *A Control Scheme for Automatic Level Crossings Under the ERTMS/ETCS Level 2/3 Operation.* IEEE Transactions on Intelligent Transportation Systems, 2017. (Cited in pages 65 and 80.)

[Good 1966] I.J. Good, I. Hacking, R.C. Jeffrey et H. Tornebohm. The estimation of probabilities: An essay on modern bayesian methods. MIT Press, 1966. (Cited in page 137.)

[Green 1967] D. M. Green et R. D. Luce. *Detection of auditory signals presented at random times.* Attention, Perception, & Psychophysics, vol. 2, no. 10, pages 441–450, 1967. (Cited in pages 60 and 78.)

[Guikema 2012] S. D. Guikema et S. M. Quiring. *Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data.* Reliability Engineering & System Safety, vol. 99, pages 178–182, 2012. (Cited in pages 20 and 88.)

[Hanley 1982] J.A. Hanley et B.J. McNeil. *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, vol. 143, no. 1, pages 29–36, 1982. (Cited in page 140.)

[Hao 2013] W. Hao et J. Daniel. *Severity of injuries to motor vehicle drivers at highway-rail grade crossings in the United States.* Transportation Research Record: Journal of the Transportation Research Board, no. 2384, pages 102–108, 2013. (Cited in pages 15 and 88.)

[Heckerman 1995] D. Heckerman, D. Geiger et D.M. Chickering. *Learning Bayesian networks: The combination of knowledge and statistical data.* Machine learning, vol. 20, no. 3, pages 197–243, 1995. (Cited in page 137.)

[Heredia-Zavoni 2012] E. Heredia-Zavoni, R. Montes-Iturrizaga, M.H. Faber et D. Straub. *Risk assessment for structural design criteria of FPSO systems. Part II: Consequence models and applications to determination of target reliabilities.* Marine Structures, vol. 28, no. 1, pages 50–66, 2012. (Cited in page 35.)

[Hu 2010] S. R. Hu, C. S. Li et C. K. Lee. *Investigation of key factors for accident severity at railroad grade crossings by using a logit model.* Safety science, vol. 48, no. 2,

pages 186–194, 2010. (Cited in pages 34 and 88.)

[Jarque 1980] C. M. Jarque et A. K. Bera. *Efficient tests for normality, homoscedasticity and serial independence of regression residuals.* Economics letters, vol. 6, no. 3, pages 255–259, 1980. (Cited in page 94.)

[Jensen 1996] F.V. Jensen. An introduction to bayesian networks, volume 210. UCL press London, 1996. (Cited in page 130.)

[Jia 2007] M. T. Jia. Accident analysis and prediction at level crossings. M.S. thesis, Beijing Jiaotong University, China., 2007. (Cited in pages 14 and 40.)

[Kallberg 2017] V.-P. Kallberg et A. Silla. *Prevention of railway trespassing by automatic sound warning-A pilot study.* Traffic injury prevention, vol. 18, no. 3, pages 330–335, 2017. (Cited in page 4.)

[Khattak 2009] A. Khattak. *Comparison of driver behavior at highway-railroad crossings in two cities.* Transportation Research Record: Journal of the Transportation Research Board, no. 2122, pages 72–77, 2009. (Cited in page 15.)

[Khattak 2012] A. Khattak, A. Sharma et Z. Luo. *Implications of using annual average daily traffic in highway-rail grade crossing safety models.* Transportation Research Board 91st Annual Meeting, No. 12-2549, 2012. (Cited in pages 22 and 88.)

[Kjærulff 2000] U. Kjærulff et Linda C. van der G. *Making sensitivity analysis computationally efficient.* Dans Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pages 317–325. Morgan Kaufmann Publishers Inc., 2000. (Cited in page 138.)

[Koiter 2006] J.R. Koiter. *Visualizing inference in Bayesian networks (M.Sc. thesis).* Delft University of Technology, 2006. (Cited in page 159.)

[Langseth 2007] H. Langseth et L. Portinale. *Bayesian networks in reliability.* Reliability Engineering & System Safety, vol. 92, no. 1, pages 92–108, 2007. (Cited in page 34.)

[Larue 2015] G. S. Larue, A. Rakotonirainy, N. L. Haworth et M. Darvell. *Assessing driver acceptance of Intelligent Transport Systems in the context of railway level crossings.* Transportation Research Part F: Traffic Psychology and Behaviour, vol. 30, pages 1–13, 2015. (Cited in pages 16 and 17.)

[Lauria 2006] E.J.M. Lauria et P.J. Duchessi. *A Bayesian belief network for IT implementation decision support.* Decision Support Systems, vol. 42, no. 3, pages 1573–1588, 2006. (Cited in page 34.)

[Lauritzen 1988] S.L. Lauritzen et D.J. Spiegelhalter. *Local computations with probabilities on graphical structures and their application to expert systems.* Journal of the Royal Statistical Society. Series B (Methodological), pages 157–224, 1988. (Cited in page 133.)

[Lemmer 2014] J.F. Lemmer et L.N. Kanal. *Propagating uncertainty in Bayesian networks by probabilistic logic sampling.* Dans Uncertainty in artificial intelligence, volume 2, page 149, 2014. (Cited in page 133.)

[Lenné 2011] M. G. Lenné, C. M. Rudin-Brown, J. Navarro, J. Edquist, M. Trotter et N. Tomasevic. *Driver behaviour at rail level crossings: Responses to flashing lights, traffic signals and stop signs in simulated rural driving.* Applied ergonomics, vol. 42, no. 4, pages 548–554, 2011. (Cited in pages 15, 16 and 55.)

[Leveson 2011] N. G. Leveson. Engineering a safer world: Systems thinking applied to safety. MIT Press, Cambridge, 2011. (Cited in page 15.)

[Liang 2016] C. Liang, M. Ghazel, E. M. El-Koursi et O. Cazier. *Statistical Analysis of Collisions at French Level Crossings.* Proceedings of the Third International Conference on Railway Technology: Research, Development and Maintenance - RAILWAYS 2016, no. 71, 2016. (Cited in page 39.)

[Liang 2017a] C. Liang, M. Ghazel, O. Cazier, L. Bouillaut et E. M. El-Koursi. *Bayesian network modeling applied on railway level crossing safety.* Dans International Conference on Reliability, Safety and Security of Railway Systems, pages 116–130. Springer, Pistoia, Italy, 2017. (Cited in page 128.)

[Liang 2017b] C. Liang, M. Ghazel, O. Cazier et E. M. El-Koursi. *A new insight on the risky behavior of motorists at railway level crossings: an observational field study.* Accident Analysis & Prevention, vol. 108, pages 181–188, 2017. (Cited in page 54.)

[Liang 2017c] C. Liang, M. Ghazel, O. Cazier et E. M. El-Koursi. *Risk analysis on level crossings using a causal Bayesian network based approach.* Transportation Research Procedia, vol. 25, pages 2172–2186, 2017. (Cited in pages 6, 22, 93, 128 and 143.)

[Liang 2018a] C. Liang, M. Ghazel et O. Cazier. *Using Bayesian networks for the purpose of risk analysis at railway level crossings.* Dans 15th IFAC Symposium on Control in Transportation Systems - CTS 2018. Springer, Savona, Italy, 2018. (Cited in page 128.)

[Liang 2018b] C. Liang, M. Ghazel, O. Cazier et L. Bouillaut. *Advanced model-based risk reasoning on automatic Level crossings.* IEEE Trans. on Intelligent Transportation Systems, 2018. (Cited in page 128.)

[Liang 2018c] C. Liang, M. Ghazel, O. Cazier et E. M. El-Koursi. *Analyzing risky behavior of motorists during the closure cycle of railway level crossings.* Safety Science, https://doi.org/10.1016/j.ssci.2017.12.008, 2018. (Cited in page 54.)

[Liang 2018d] C. Liang, M. Ghazel, O. Cazier et E. M. El-Koursi. *Developing accident prediction model for railway level crossings.* Safety Science, vol. 101, pages 48–59, 2018. (Cited in pages 86, 93, 110 and 143.)

[Liang 2018e] C. Liang, M. Ghazel, O. Cazier et E. M. El-Koursi. *A risk assessment study on accidents at French level crossings using Bayesian belief networks.* International Journal of Injury Control and Safety Promotion, https://doi.org/10.1080/17457300.2017.1416480, 2018. (Cited in page 128.)

[Liu 2016] B. Liu, M. Ghazel et A. Toguyéni. *Model-based diagnosis of multi-track level crossing plants.* IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 2, pages 546–556, 2016. (Cited in page 4.)

[Lord 2010] D. Lord et F. Mannering. *The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives.* Transportation Research

Part A: Policy and Practice, vol. 44, no. 5, pages 291–305, 2010. (Cited in pages 20, 21, 88, 98 and 99.)

[Lu 2016] P. Lu et D. Tolliver. *Accident prediction model for public highway-rail grade crossings.* Accident Analysis & Prevention, vol. 90, pages 73–81, 2016. (Cited in pages 22, 89, 99, 107, 110 and 119.)

[Madsen 2004] K. Madsen, H. B. Nielsen et O. Tingleff. *Methods for non-linear least squares problems (2nd ed.).* Informatics and Mathematical Modelling, Technical University of Denmark, 2004. (Cited in page 96.)

[Mahboob 2013] Qamar Mahboob. *A Bayesian Network methodology for railway risk, safety and decision support.* Technische Universitat Dresden, 2013. (Cited in pages 25 and 26.)

[Malyshkina 2009] N.V. Malyshkina et F.L. Mannering. *Markov switching multinomial logit model: an application to accident-injury severities.* Accident Analysis & Prevention, vol. 41, no. 4, pages 829–838, 2009. (Cited in pages 30 and 129.)

[McCollister 2007] G. M. McCollister et C. C. Pflaum. *A model to predict the probability of highway rail crossing accidents.* Proceedings of the Institution of Mechanical Engineers, Part F: Journal of rail and rapid transit, vol. 221, no. 3, pages 321–329, 2007. (Cited in pages 15 and 40.)

[Medina 2015] J. C. Medina et R. F. Benekohal. *Macroscopic models for accident prediction at railroad grade crossings: comparisons with US Department of Transportation accident prediction formula.* Transportation Research Record: Journal of the Transportation Research Board, no. 2476, pages 85–93, 2015. (Cited in pages 22 and 88.)

[Mekki 2012] A. Mekki, M. Ghazel et A. Toguyeni. *Validation of a new functional design of automatic protection systems at level crossings with model-checking techniques.* IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 2, pages 714–723, 2012. (Cited in page 4.)

[Miaou 1994] S. P. Miaou. *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions.* Accident Analysis & Prevention, vol. 26, no. 4, pages 471–482, 1994. (Cited in pages 21, 96 and 98.)

[Miranda-Moreno 2005] L. Miranda-Moreno, L. Fu, F. F. Saccomanno et A. Labbe. *Alternative risk models for ranking locations for safety improvement.* Transportation Research Record: Journal of the Transportation Research Board, no. 1908, pages 1–8, 2005. (Cited in pages 22, 89 and 119.)

[Murphy 2002] K.P. Murphy et S. Russell. *Dynamic bayesian networks: representation, inference and learning.* 2002. (Cited in page 168.)

[Nadkarni 2004] S. Nadkarni et P. P. Shenoy. *A causal mapping approach to constructing Bayesian networks.* Decision support systems, vol. 38, no. 2, pages 259–281, 2004. (Cited in page 34.)

[Niittymaki 1998] J. Niittymaki et S. Kikuchi. *Application of fuzzy logic to the control of a pedestrian crossing signal.* Transportation Research Record: Journal of the Transportation Research Board, no. 1651, pages 30–38, 1998. (Cited in pages 32 and 129.)

[Ogden 2007] B. D. Ogden. *Railroad-highway grade crossing handbook: Revised.* Rapport technique, Tech. Rep. FHWA-SA-07-010, US Department of Transportation, Federal Highway Administration, Washington, DC, 2007. (Cited in page 17.)

[Oh 2006] J. Oh, S. P. Washington et D. Nam. *Accident prediction model for railway-highway interfaces.* Accident Analysis & Prevention, vol. 38, no. 2, pages 346–356, 2006. (Cited in pages 17, 22, 87, 88, 99 and 119.)

[Pearl 1986] J. Pearl. *Fusion, propagation, and structuring in belief networks.* Artificial intelligence, vol. 29, no. 3, pages 241–288, 1986. (Cited in page 133.)

[Pearl 1995] J. Pearl et T.S. Verma. *A theory of inferred causation.* Studies in Logic and the Foundations of Mathematics, vol. 134, pages 789–811, 1995. (Cited in page 142.)

[Pearson 1900] K. Pearson. *X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 50, no. 302, pages 157–175, 1900. (Cited in page 106.)

[Peterson 1981] J.L. Peterson. *Petri net theory and the modeling of systems.* 1981. (Cited in page 30.)

[Plesse 2017] G. Plesse. *Des détecteurs d'obstacles déployés aux passages à niveau.* Rapport technique, France. From http://www.leparisien.fr/info-paris-ile-de-france-oise/transports/des-detecteurs-d-obstacles-deployes-aux-passages-a-niveau-02-06-2017-7011714.php, 2017. (Cited in pages 5 and 40.)

[Powers 2011] D. M. Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* Journal of Machine Learning Technologies, vol. 2, no. 1, pages 37–63, 2011. (Cited in page 140.)

[Read 2016] G. J. M. Read, P. M. Salmon, M. G. Lenné et N. A. Stanton. *Walking the line: understanding pedestrian behaviour and risk at rail level crossings with cognitive work analysis.* Applied ergonomics, vol. 53, pages 209–227, 2016. (Cited in pages 4, 15, 16 and 17.)

[Reed 2002] G. F. Reed, F. Lynn et B. D. Meade. *Use of coefficient of variation in assessing variability of quantitative assays.* Clinical and diagnostic laboratory immunology, vol. 9, no. 6, pages 1235–1239, 2002. (Cited in page 145.)

[Richards 1990] S. H. Richards et K. W. Heathington. *Assessment of warning time needs at railroad-highway grade crossings with active traffic control.* Transportation Research Record, vol. 1254, pages 72–84, 1990. (Cited in page 14.)

[Ridout 2001] M. Ridout, J. Hinde et C. DeméAtrio. *A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives.* Biometrics, vol. 57, no. 1, pages 219–223, 2001. (Cited in pages 21 and 98.)

[Rudin-Brown 2012] C. M. Rudin-Brown, M. G. Lenné, J. Edquist et J. Navarro. *Effectiveness of traffic light vs. boom barrier controls at road–rail level crossings: A simulator study.* Accident Analysis & Prevention, vol. 45, pages 187–194, 2012. (Cited in page 16.)

[Saccomanno 2003] F. F. Saccomanno, C. G. Ren et L. Fu. *Collision prediction models for highway-rail grade crossings in Canada.* Dans 82nd Annual Meeting of the Transportation Research Board, at Washington, DC, 2003. (Cited in page 88.)

[Saccomanno 2004] F. F. Saccomanno, L. Fu et L. Miranda-Moreno. *Risk-based model for identifying highway-rail grade crossing blackspots.* Transportation Research Record: Journal of the Transportation Research Board, no. 1862, pages 127–135, 2004. (Cited in page 119.)

[Salmon 2013] P. M. Salmon, G. J. Read, N. A. Stanton et M. G. Lenné. *The crash at Kerang: Investigating systemic and psychological factors leading to unintentional non-compliance at rail level crossings.* Accident Analysis & Prevention, vol. 50, pages 1278–1288, 2013. (Cited in page 15.)

[Salmon 2016] P. M. Salmon, M. G. Lenné, G. J. Read, C. M. Mulvihill, M. Cornelissen, G. H. Walker, K. L. Young, N. Stevens et N. A. Stanton. *More than meets the eye: using cognitive work analysis to identify design requirements for future rail level crossing systems.* Applied ergonomics, vol. 53, pages 312–322, 2016. (Cited in page 15.)

[San Kim 2013] D. San Kim et W. C. Yoon. *An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK.* Safety science, vol. 60, pages 57–68, 2013. (Cited in pages 15 and 55.)

[Schoppert 1968] D. W. Schoppert, W. Dan et M. Alan. Factors influencing safety at highway-rail grade crossings (nchrp report 50). NAS-NRC, 1968. (Cited in page 19.)

[Seise 2010] A. Seise, V.-P. Kallberg et A. Silla. *The effect of speed bumps on driving speeds at road-railway level crossings.* Dans Proc. 11th Level Crossing Symp. Tokyo, pages 8–17, 2010. (Cited in page 4.)

[Shachter 2013] R. D. Shachter et M. A. Peot. *Simulation approaches to general probabilistic inference on belief networks.* arXiv preprint arXiv:1304.1526, 2013. (Cited in page 133.)

[Shappel 2000] S. A. Shappel et D. A. Wiegmann. *The human factors analysis and classification system–HFACS (No. DOT/FAA/AM-00/7).* US Federal Aviation Administration, Office of Aviation Medicine, 2000. (Cited in pages 15 and 55.)

[Silla 2012] A. Silla et J. Luoma. *Main characteristics of train-pedestrian fatalities on Finnish railroads.* Accident Analysis & Prevention, vol. 45, no. 1, pages 61–66, 2012. (Cited in page 4.)

[Silmon 2010] J. Silmon et C. Roberts. *Using functional analysis to determine the requirements for changes to critical systems: Railway level crossing case study.* Reliability Engineering & System Safety, vol. 95, no. 3, pages 216–225, 2010. (Cited in pages 4 and 82.)

[SNCF Réseau 2010] SNCF. SNCF Réseau. *Statistical analysis of accidents at LXs.* Rapport technique, France, 2010. (Cited in page 94.)

[SNCF Réseau 2011] SNCF. SNCF Réseau. *World Conference of Road Safety at Level Crossings.* Rapport technique, France. From http://www.planetoscope.com/automobile/1271-nombre-de-collisions-aux-passages-a-niveau-en-france.html, 2011. (Cited in page 4.)

[SNCF 2015] SNCF. *Research on the material of level crossing in 2014.* Rapport technique, France, 2015. (Cited in page 5.)

[SNCF 2017] SNCF. *MORIPAN Report v8.* Rapport technique, France, 2017. (Cited in pages 80 and 168.)

[Špačková 2013] Olga Špačková et Daniel Straub. *Dynamic Bayesian network for probabilistic modeling of tunnel excavation processes.* Computer-Aided Civil and Infrastructure Engineering, vol. 28, no. 1, pages 1–21, 2013. (Cited in page 35.)

[Stanton 2011] N. A. Stanton et G. H. Walker. *Exploring the psychological factors involved in the Ladbroke Grove rail accident.* Accident Analysis & Prevention, vol. 43, no. 3, pages 1117–1127, 2011. (Cited in pages 15 and 17.)

[Stefanova 2015] T. Stefanova, J. M. Burkhardt, A. Filtness, C. Wullems, A. Rakotonirainy et P. Delhomme. *Systems-based approach to investigate unsafe pedestrian behaviour at level crossings.* Accident Analysis & Prevention, vol. 81, pages 167–186, 2015. (Cited in pages 16 and 17.)

[Taylor 1967] M. M. Taylor, P. H. Lindsay et S. M. Forbes. *Quantification of shared capacity processing in auditory and visual discrimination.* Acta Psychologica, vol. 27, pages 223–229, 1967. (Cited in pages 60 and 78.)

[Taylor 2008] K. Taylor. *Addressing road user behavioural changes at railway level crossings.* Dans ACRS-Travelsafe National Conference, pages 368–375, 2008. (Cited in page 82.)

[Tey 2011] L. S. Tey, L. Ferreira et A. Wallace. *Measuring driver responses at railway level crossings.* Accident Analysis & Prevention, vol. 43, no. 6, pages 2134–2141, 2011. (Cited in pages 16 and 17.)

[Utkin 2015] L. V. Utkin, F. P. A. Coolen et S. V. Gurov. *Imprecise inference for warranty contract analysis.* Reliability Engineering & System Safety, vol. 138, pages 31–39, 2015. (Cited in pages 21 and 98.)

[Vatn 2002] J. Vatn et H.A. Svee. *Risk Based Approach to Determine Ultrasonic Inspection Frequencies in Railway Applications.* Dans Proceedings of the 22nd ESReDA Seminar, Madrid, Spain May, pages 27–28, 2002. (Cited in page 169.)

[Ward 1995] N. J. Ward et G. J. Wilde. *Field observation of advance warning/advisory signage for passive railway crossings with restricted lateral sightline visibility: an experimental investigation.* Accident Analysis & Prevention, vol. 27, no. 2, pages 185–197, 1995. (Cited in page 14.)

[Weakliem 1999] D. L. Weakliem. *A critique of the Bayesian information criterion for model selection.* Sociological Methods & Research, vol. 27, no. 3, pages 359–397,

1999. (Cited in page 106.)

[Weber 2012] P. Weber, G. Medina-Oliva, C. Simon et B. Iung. *Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas.* Engineering Applications of Artificial Intelligence, vol. 25, no. 4, pages 671–682, 2012. (Cited in pages 10, 35 and 129.)

[Welford 1952] A. T. Welford. *The 'psychological refractory period' and the timing of high-speed performance—a review and a theory.* British Journal of Psychology, vol. 43, no. 1, pages 2–19, 1952. (Cited in pages 60 and 78.)

[Wigglesworth 2001] E. C. Wigglesworth. *A human factors commentary on innovations at railroad–highway grade crossings in Australia.* Journal of Safety Research, vol. 32, no. 3, pages 309–321, 2001. (Cited in pages 16 and 17.)

[Willsher 2017] K. Willsher. *Children killed as train and school bus collide in southern France.* Rapport technique, France. http://www.theguardian.com/world/2017/dec/14/school-bus-and-train-in-serious-accident-in-southern-france, 2017. (Cited in page 14.)

[Wilson 2005] J. R. Wilson et B. J. Norris. *Rail human factors: Past, present and future.* Applied ergonomics, vol. 36, no. 6, pages 649–660, 2005. (Cited in pages 15 and 55.)

[Wilson 2014] J. R. Wilson. *Fundamentals of systems ergonomics/human factors.* Applied ergonomics, vol. 45, no. 1, pages 5–13, 2014. (Cited in pages 15 and 55.)

[Woods 2008] M. D. Woods, R. Slovak, E. Schnieder et E. M. El-Koursi. *Safer European Level Crossing Appraisal and Technology (SELCAT)-D3 Report on Risk Modeling Techniques for level crossing risk and system safety evaluation.* Rapport technique, Rail Safety and Standards Board, 2008. (Cited in page 19.)

[Wullems 2011] C. Wullems. *Towards the adoption of low-cost rail level crossing warning devices in regional areas of Australia: A review of current technologies and reliability issues.* Safety science, vol. 49, no. 8, pages 1059–1073, 2011. (Cited in page 82.)

[Yan 2003] L. Yan, R. H. Dodier, M. L. Mozer et R. H. Wolniewicz. *Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic.* Dans Proceedings of the 20th International Conference on Machine Learning (ICML-03), pages 848–855, 2003. (Cited in page 140.)

[Young 2015] K. L. Young, M. G. Lenné, V. Beanland, P. M. Salmon et N. A. Stanton. *Where do novice and experienced drivers direct their attention on approach to urban rail level crossings?* Accident Analysis & Prevention, vol. 77, pages 1–11, 2015. (Cited in page 16.)

[Yuan 2002] C. Yuan et M.J. Druzdzel. *An importance sampling algorithm based on evidence pre-propagation.* Dans Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, pages 624–631. Morgan Kaufmann Publishers Inc., 2002. (Cited in page 133.)

Appendix A

Reference Materials

A.1 Predictive accuracy comparison between λ_{10Y} and λ_{10P}

LL test results are shown in Table A.1. One can notice that for λ_{10Y} model combined with either the Poisson or NB distribution, its GOFs are significantly better than λ_{10P} model's GOFs according to LL results. Furthermore, the GOF of λ_{10Y} combined with the NB distribution is better than when combined with the Poisson distribution.

Table A.1. Model quality comparison between λ_{10Y} and λ_{10P}

Parameter	**λ_{10Y} Poisson**	**λ_{10Y} NB**	**λ_{10P} Poisson**	**λ_{10P} NB**
Railway traffic characteristics				
Average daily railway traffic	×	×	×	×
Railway speed limit	×	×		
Roadway traffic characteristics				
Average daily road traffic	×	×	×	×
Annual road accidents	×	×	×	×
LX characteristics				
Alignment	×	×		
Profile	×	×		
LX width	×	×		
Crossing length	×	×		
Region	×	×		
AIC	-190,744 (1)	-190,744 (1)	-190,591 (2)	-190,591 (2)
BIC	-190,670 (1)	-190,670 (1)	-190,573 (2)	-190,573 (2)
PCS	65,796 (1)	65,796 (1)	53,108 (2)	53,108 (2)
DF	83,313	83,313	83,319	83,319
LL	-2,599 (2)	-2,596 (1)	-2,631 (4)	-2,629 (3)
Goodness score (the lower, the better)	5	4	10	9

Based on the predicted probability of the accident frequency observed, further Cumulative Distribution Function (CDF) analysis with regard to the Poisson and the NB distributions is performed to evaluate the quality of the accident frequency prediction model combined with these two statistical distributions. As shown in Fig. A.1, the relationship between the CDF and the corresponding probability of a given event is depicted. $\hat{P}(\bullet)$ denotes the predicted probability of a given event obtained through the Poisson or NB distribution; O_i is the observed accident frequency and λ_i is the estimated accident

frequency. The blue curve "CDF NB λ_{10P}, $O_i > \lambda_i$" represents the CDF of event " $O_i > \lambda_i$" obtained through the NB distribution combined with the λ_{10P}; the red curve "CDF NB λ_{10P}, $O_i <= \lambda_i$" represents the CDF of event " $O_i <= \lambda_i$" obtained through the NB distribution combined with the λ_{10P}; the green curve "CDF POI λ_{10P}, $O_i > \lambda_i$" represents the CDF of event " $O_i > \lambda_i$" obtained through the Poisson distribution combined with the λ_{10P}; the violet curve "CDF POI λ_{10P}, $O_i <= \lambda_i$" represents the CDF of event " $O_i <= \lambda_i$" obtained through the Poisson distribution combined with the λ_{10P}. The interpretation of the remaining curves involving the λ_{10Y} can be similarly obtained. Given that some curves are almost covered by some others in Fig. A.1, the extracted results of CDF analysis shown in Table A.2 become clearer for discussion.

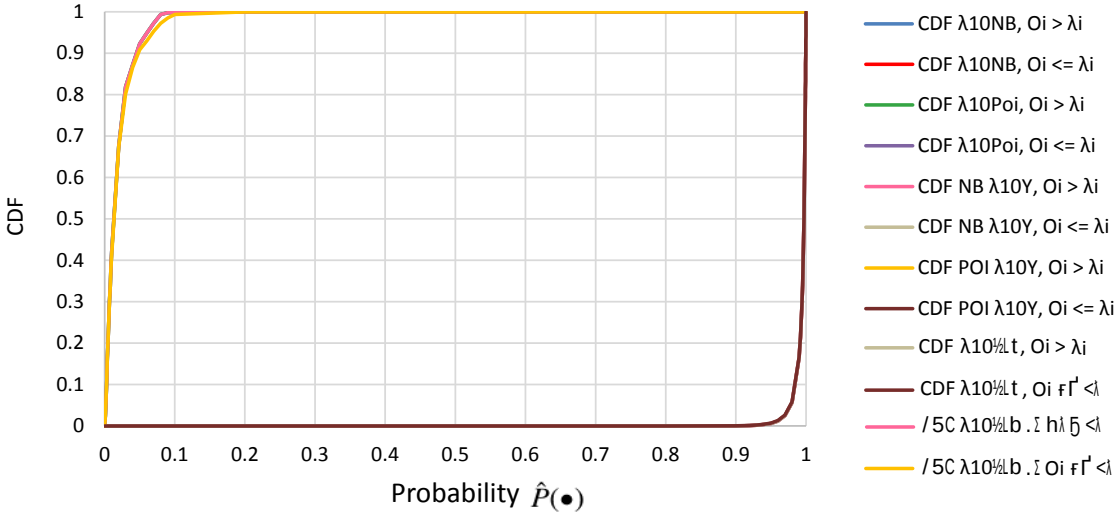


Fig. A.1. CDF of the Poisson and the NB distributions combined with the λ_{10P} and λ_{10Y} models according to the estimated probability

Table A.2. The extracted results of CDF analysis

Model CDF	$\hat{P}(O_i > \lambda_i) >$ 0.005 (CDF in percent)	$\hat{P}(O_i > \lambda_i) >$ 0.05 (CDF in percent)	$\hat{P}(O_i <= \lambda_i) >$ 0.95 (CDF in percent)	$\hat{P}(O_i <= \lambda_i) >$ 0.995 (CDF in percent)
CDF NB λ_{10P}	85.29 (3)	6.61 (1)	99.62 (1)	57.19 (3)
CDF NB λ_{10Y}	79.10 (2)	7.68 (3)	99.36 (3)	66.07 (1)
CDF POI λ_{10P}	85.29 (3)	6.82 (2)	99.61 (2)	57.15 (4)
CDF POI λ_{10Y}	78.89 (1)	9.17 (4)	99.27 (4)	65.94 (2)

Table A.2 indicates that:

1) CDF NB λ_{10P}, $O_i > \lambda_i$:
In 85.29% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 6.61% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

2) CDF POI λ_{10P}, $O_i > \lambda_i$:
In 85.29% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 6.82% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

3) CDF NB λ_{10Y}, $O_i > \lambda_i$:
In 79.10% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 7.68% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

4) CDF POI λ_{10Y}, $O_i > \lambda_i$:
In 78.89% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.005; in 9.17% of cases, $\hat{P}(O_i > \lambda_i)$ is more than 0.05;

5) CDF NB λ_{10P}, $O_i <= \lambda_i$:
In 99.62% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 57.19% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

6) CDF POI λ_{10P}, $O_i <= \lambda_i$:
In 99.61% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 57.15% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

7) CDF NB λ_{10Y}, $O_i <= \lambda_i$:
In 99.36% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 66.07% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

8) CDF POI λ_{10Y}, $O_i <= \lambda_i$:
In 99.27% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.95; in 65.94% of cases, $\hat{P}(O_i <= \lambda_i)$ is more than 0.995;

According to the CDF analysis results shown in Table A.2, in the cases of "$\hat{P}(O_i > \lambda_i) > 0.005$" and "$\hat{P}(O_i <= \lambda_i) > 0.995$", for the λ_{10Y} model combined with either the Poisson or the NB distribution, its GOFs are significantly better than λ_{10P} model's GOFs. In the cases of "$\hat{P}(O_i > \lambda_i) > 0.05$" and "$\hat{P}(O_i <= \lambda_i) > 0.95$", the criteria of the two models combined with the Poisson and the NB distributions have no obvious distinction, in particular, for the criterion "$\hat{P}(O_i <= \lambda_i) > 0.95$". Furthermore, λ_{10Y} combined with the NB distribution shows a slightly better quality than when combined with the Poisson distribution.

As shown in Table A.4, f_k denotes the percentage of samples of observed annual accident frequency with k accidents involved in a given year (f_k = the number of samples of observed annual accident frequency involving k accidents occurring in a given year / the total number of samples n). The estimated relative annual accident frequency reflected by estimated probabilities on average is computed as: $\hat{f}_k = \sum_{i=1}^{n} \hat{P}(X_i = k)/n$, where $\hat{P}(X_i = k)$ is the estimated probability of k accidents occurring at a given SAL2 in a given year. According to the goodness of predictive accuracy ranked in brackets, the NB distribution shows a higher predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during the 10-year period, particularly, when combining with the λ_{10Y}. In the cases of 0, 1 and 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the NB distribution combined with the λ_{10Y} takes the first place in all the cases, which means that the probabilities of accident occurrence predicted by the NB distribution combined with the λ_{10Y} are closest to the actual frequencies of accident occurrence. In the case of more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the NB distribution combined with the λ_{10Y} takes the second place, with the deviation of only 0.0002% compared with f_k, the actual percentage of observed annual accident frequency samples. In fact, there are no SAL2 LXs showing more than 2 accidents in the same year during the 10-year period considered.

Table A.3. The predictive accuracy comparison between the Poisson distribution and the NB distribution

# Annual accidents considered (k)	Observed annual accident frequency (f_k in percent)	NB-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{10Y} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	NB-λ_{10P} estimated relative annual accident frequency ($\hat{f}_k$ in percent)	POI-λ_{10P} estimated relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.4371	99.3915 (1)	99.3903 (2)	99.3279 (3)	99.3255 (4)
1	0.5485	0.5999 (1)	0.6033 (2)	0.6647 (3)	0.6673 (4)
2	0.0144	0.0077 (1)	0.0062 (3)	0.0069 (2)	0.0055 (4)
> 2	0	0.0002 (2)	0.0001 (1)	0.0001 (1)	0.0001 (1)
Goodness score (the lower, the better)		5	8	9	13

A.2 Values of Parameters in Statistical Accident Prediction Model

In this chapter, we give the detailed values of "Region risk factor" defined in 5.2.

Table A.4. Detailed values of "Region risk factor"

Region	Region risk factor
Alsace	0.5118
Aquitaine	0.3523
Auvergne	0.2442
Basse-Normandie	0.2143
Bourgogne	0.2755
Bretagne	0.3239
Centre-Val de Loire	0.2669
Champagne-Ardenne	0.3426
Franche-Comté	0.3676
Haute-Normandie	0.2036
Îl-de-France	0.7747
Languedoc-Roussillon	0.5951
Limousin	0.1739
Lorraine	0.4317
Midi-Pyrénées	0.3187
Nord-Pas-de-Calais	0.4165
Pays de la Loire	0.3156
Picardie	0.3755
Poitou-Charentes	0.2171
Provence-Alpes-Côte-d'Azur	0.5902
Rhône-Alpes	0.4694