Generating a Risk Profile for Car Insurance Policyholders: A Deep Learning Conceptual Model

Mohammad Siami
Decision Systems and e-Service Intelligence Laboratory,
Centre for Artificial Intelligence (CAI), Faculty of Engineering and IT,
University of Technology Sydney (UTS)
PO Box 123, Broadway, NSW 2007, Australia
Email: Mohammad.SiamiNamini@uts.edu.au

Mohsen Naderpour
Decision Systems and e-Service Intelligence Laboratory,
Centre for Artificial Intelligence (CAI), Faculty of Engineering and IT,
University of Technology Sydney (UTS)
PO Box 123, Broadway, NSW 2007, Australia
Email: Mohsen.Naderpour@uts.edu.au

Jie Lu
Decision Systems and e-Service Intelligence Laboratory,
Centre for Artificial Intelligence (CAI), Faculty of Engineering and IT,
University of Technology Sydney (UTS)
PO Box 123, Broadway, NSW 2007, Australia
Email: Jie.Lu@uts.edu.au

Abstract

In recent years, technological improvements have provided a variety of new opportunities for insurance companies to adopt telematics devices in line with usage-based insurance models. This paper sheds new light on the application of big data analytics for car insurance companies that may help to estimate the risks associated with individual policyholders based on complex driving patterns. We propose a conceptual framework that describes the structural design of a risk predictor model for insurance customers and combines the value of telematics data with deep learning algorithms. The model's components consist of data transformation, criteria mining, risk modelling, driving style detection, and risk prediction. The expected outcome is our methodology that generates more accurate results than other methods in this area.

Keywords

Risk assessment, usage-based insurance, deep learning, telematics

1 Introduction

The insurance market is very competitive and, as such, many insurance companies would rather provide products and services to their clients based on a personal risk score. However, to do this, they need access to dynamic risk profiles for all their policyholders to be able to manage their own organisational exposure and capital adequacy rates. There are many different approaches to risk calculation. Vehicle- or driver-specific parameters and socio-demographic variables have been used for many years. Traditionally, customer claims and their historical records have been the most valuable data for calculating a customer's risk. These methods are widely accepted; however, they are fraught with significant challenges. One is that they cannot be used to assess the exposure of policyholders based on their driving behaviour. Recently, insurance companies in Australia, the USA, and Canada have used telematics and in-vehicle data records to provide consistent services to low-risk customers. In fact, the emergence of telematics has introduced an entirely new area for insurers to estimate the risk of their clients based on their daily driving behaviours (Baecke and Bocca 2017).

The application of telematics data in the insurance sector also provides some scientific and commercial opportunities for insurance service providers. Telematics devices could help insurers study car movement parameters, such as location, speed, and acceleration, and how these parameters change under various conditions, which could reflect driver habits (Wahlström et al. 2015). These devices are certainly generating large datasets and creating an overwhelming dilemma for many insurance companies. In response to these emerging issues, our major three research questions are: 1) How can telematics data be used to determine the driving signature of individual policyholders? 2) How can similar patterns in different driving habits be extracted? And, in this regard, how can the characteristics of short trips, long trips, and highways trips be determined? 3) How can abnormalities in driving characteristics within these telematics datasets be detected?

Answering these research questions could help to generate a dynamic risk profile for customers. For many years, risk assessment directly addressed by researchers in financial risk prediction (Siami et al. 2011; Siami et al. 2014), situation awareness (Naderpour et al. 2014a; Naderpour et al. 2014b), and safety assessment (Purba et al. 2014). But, so far, a limited amount of literature has been published on risk assessment and data analytics with telematics for usage-based insurance. This prospective study has been designed to investigate the use of telematics data in the insurance market using deep neural networks models. We propose a conceptual model that generates a dynamic risk profile for insurance customers based on complex driving patterns. This paper describes our framework and investigates the role of deep learning in the context of risk prediction models.

The rest of this paper is organised as follows. Section 2 provides a brief overview of telematics and its applications in insurance and reviews some of the literature on this topic. Section 3 presents the conceptual model and its components. Section 4 presents the implementation process. Finally, Section 5 concludes the paper and describes future work.

2 Literature Review

A telematics device is a kind of in-vehicle data recorder that includes a GPS sensor with the ability to transmit data to a remote server. It is a hardware device that can be incorporated into a vehicle and can record the characteristics of driving habits. Recently, these devices have been adopted in the car insurance industry to help insurers track the behaviour of policyholders and generate a dynamic risk profile prior to making a claim.

Figure 1 provides an overview of telematics devices and some of the information they provide. According to Duri et al. (2002), the use of telematics devices can be categorized into three groups: location information, diagnosis and roadside assistance, and pay-for-use insurance. The first two items are beyond the scope of this paper, but the last one helps insurers provide tailored services to their customers through the data collect. Husnjak et al. (2015) explain usage-based insurance as a pricing method whereby insurers ask their customers to pay based on their driving characteristics. This includes both pay-as-you-drive (PAYD) and pay-as-how-you-drive (PHYD) schemes. These two methods share similar ideas, but they use telematics data differently. PAYD mostly focuses on the overall distance the driver drives, whereas PHYD insurance is based on how risky a driver behaves behind the wheel (Baecke and Bocca 2017).

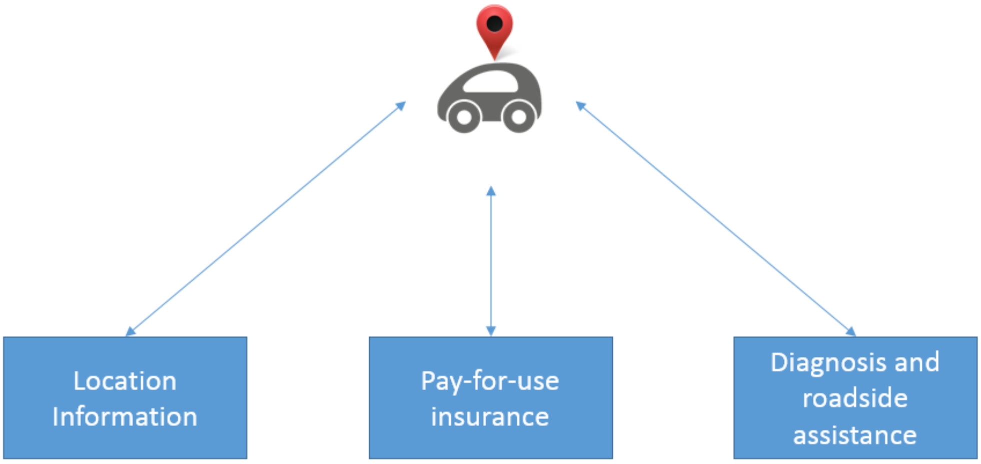

Figure 1: Telematics service provision and its application (Duri et al. 2002)

The application of telematics data in the insurance sector is new; therefore, only a few studies have investigated Big Data analytics in telematics. However, studying driving habits by applying the value of telematics data is an attractive research topic. Each driver generates a series of actions when driving. These actions are unique to each person and are repeated with some variance in different driving conditions. A driver's actions can be measured individually, and the different driving conditions might include day, night, rain or heavy traffic. For example, on a trip the driver behaves in a way that generates a sequence of behaviours that shows a short, long, fast, or slow trip. These patterns could be unique for each person and might be similar in a group of drivers with similar driving habits (Lin et al. 2014).

A few papers include a discussion about identifying driver signatures through supervised learning (Baecke and Bocca 2017; Paefgen et al. 2014). Traditional classification methods might perform well with only a limited number of drivers in the training data but, since the possibility of this is rare in the real world, these methods are not applicable. In a similar vein, if the number of classes was increased, training classifiers would become more difficult. Deep learning (LeCun et al. 2015), as a more recent approach, has shown acceptable performance in computer vision, speech recognition, and natural language processing problems (Abdel-Hamid et al. 2014; Collobert and Weston 2008; Krizhevsky et al. 2012). But the application of these algorithms in risk profile prediction has been limited, and the accuracy of current models (Dong et al. 2016; Dong et al. 2017; Huang et al. 2016) needs improvement. For example, Dong et al. (2016) proposed a new transformation method for reshaping GPS data into a new form that makes it possible for deep learning algorithms to process. Their work includes some architectures for driving signature detection and distinguishing the driving style of individuals with 1D convolutional neural networks and recurrent neural networks. Dong et al. (2017) proposed an autoencoder regularized network (ARNet) model to solve a further real world problem in insurance with a deep neural network algorithm. They estimated the correct number of drivers for each car insurance policyholder based on the historical records of a trajectory dataset. Their results indicate that their ARNet network outperforms other methods in this area. Liu et al. (2017) proposed a novel visualization method connecting driving styles to colours in an RGB colour model using a deep sparse auto encoder. This method could help us find unique driving patterns for risk prediction.

This research is planned according to the practice of the proposed research methodology by (Niu et al. 2009), which has been applied in Business Intelligence and Information systems. The primary goal of this research is to develop a risk prediction framework by applying the value of telematics data that also considers the characteristics and capabilities of deep learning. The data generated from telematics devices captures some the driving attributes of drivers, such as speed, acceleration, etc. as a time-series. Deep learning could help us exploit the value of those characteristics to identify standard or risky driving patterns and generate a risk profile for all drivers.

3 Conceptual Model

In this research, we propose a conceptual model for developing a decision support system to help car insurers evaluate the risk of their policyholders based on their driving styles, and rate them according to the behaviour of other drivers. The framework provides an opportunity for insurance companies to reduce their risk and manage the capital adequacy rates. It consists of a risk rating model and an automatic feature extractor, which identifies and classifies unique driving behaviour. Figure 2 illustrates the model; each component is explained in following subsections.

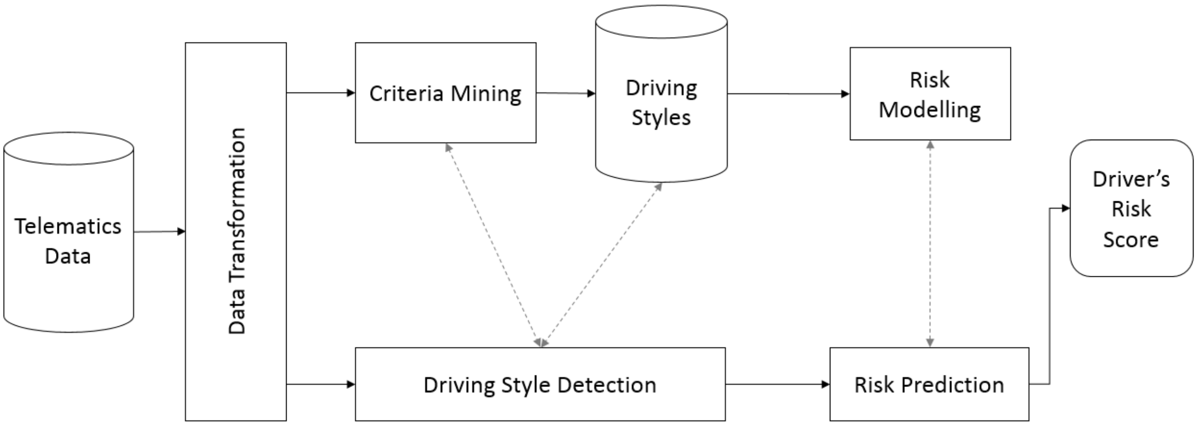

Figure 2: The conceptual model

3.1 Data Transformation

The data transformation component transforms the GPS data into a new structure that allows processing by a deep learning algorithm. Deep learning algorithms operate on complex data with many dimensions, such as images. Our model transforms the data using the methodology presented by (Dong et al. 2016).

3.2 Criteria Mining

The first step in developing our decision support system for car insurance companies is to determine the criteria for the risk profiles. Unique driving styles could be a key criterion in the risk model. Therefore, we propose a criteria miner with deep learning to generate a driving styles database. These driving behaviours might be normal or abnormal, and understanding these patterns could help inform risk prediction in the future. Figure 3 depicts the structure of criteria miner.

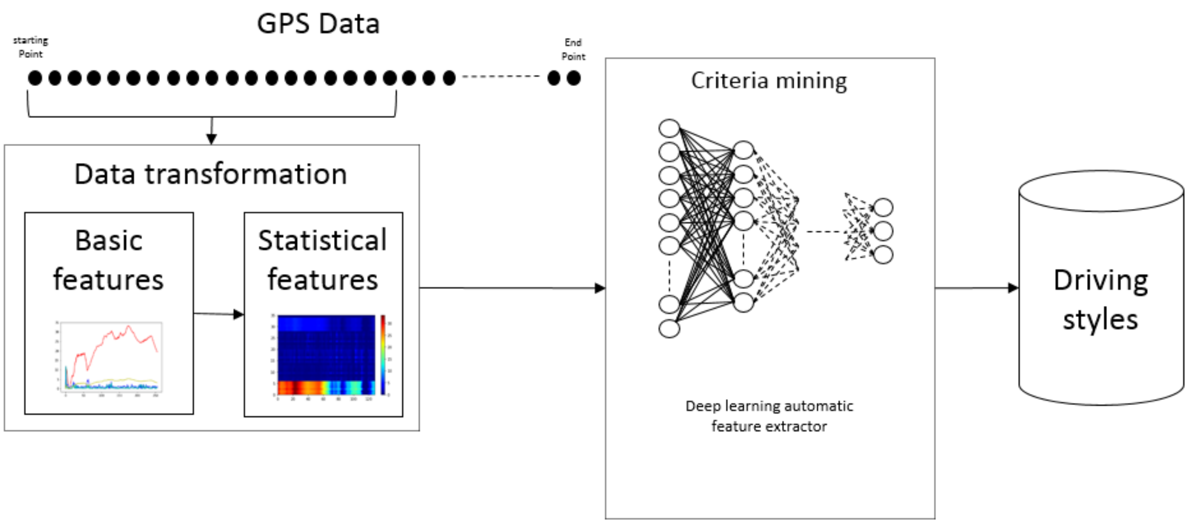

Figure 3: Criteria mining with deep learning

The criteria mining model is a deep automatic feature extractor algorithm that extracts unique patterns from the dataset. The model will rely on a deep neural network architecture, such as a deep auto encoder.

3.3 Risk Modelling

The risk modelling component generates a risk matrix with the following structure:

$$Risk\ modelling\ matrix = \begin{array}{c c} & \begin{matrix} D_1 & D_2 & \cdots & D_j \end{matrix} \\ \begin{matrix} C_1 \\ \vdots \\ C_i \end{matrix} & \begin{bmatrix} O_{11} & \cdots & O_{1j} \\ \vdots & \ddots & \vdots \\ O_{i1} & \cdots & O_{ij} \end{bmatrix} \end{array} \qquad (1)$$

where *j* is the number of drivers, *i* is the number of unique styles extracted from the criteria miner, and O_{ij} is the number of times that a particular driving style *(C_i)* is repeated by driver *j*.

We propose a new equation to calculate the risk weight of each driving style based on the assumption that most drivers will behave normally and few will drive dangerously. Hence, the probability of dangerous behaviour is lower than normal behaviour. Consequently, the following equation could find a suitable risk weight for each criterion:

$$W_{C_i} = \frac{\sum_i \sum_j o_{ij}}{\sum_j o_{ij}} \qquad (2)$$

where W_{Ci} is the risk weight of the *i*th driving style, and O_{ij} is the number of times that the *j*th driver repeats the *ith* driving style. The proposed equation assigns a high-risk weight to driving styles with a low probability of occurrence.

3.4 Driving Style Detection

This component uses the algorithms in the criteria mining component and the driving styles database to generate the risk model for the data of new drivers. The driving style detection component uses criteria miner and the driving styles database to generate the risk model for new drivers. First, the criteria miner model is used to interpret the transformed GPS data of new drivers to build the corresponding driving patterns, then driving style detection looks into the driving style database to find a similar pattern for the data of new drivers. Consequently, the output of this step is some driving patterns that exist in driving styles database and similar to the driving styles of new drivers. In this part, the most similar patterns which exist in driving styles databases are critical. Because, it might be happen some new patterns are not similar to ones in the existing styles database. In this way, we select the most similar driving characteristics that exist in our driving styles database.

3.5 Risk Prediction

The risk prediction component calculates a score for new drivers based on the similarly between their driving characteristics and our driving style database. These characteristics are extracted by driving style detection model with using criteria mining model and driving style database. The relationship between these two major components could help us to find the risk score of drivers by using the following equation:

$$(Risk\ Score)_j = \sum_i W_{C_i} * O_{ij} \qquad (3)$$

where W_{Ci} is the risk weight of the *ith* driving style, O_{ij} is the number of times that the *jth* driver repeats the *ith* driving style. The final outcome of our proposed framework is a relative score which is applicable for comparing the risk of each driver with others based on their driving characteristics and our knowledge based system. This score helps insurer to compare the risk level of a particular driver with a benchmark policy holder that they know the risk level of them.

4 Implementation

This section describes the implementation steps and expected outputs for each component, followed by a brief explanation of how the model will be validated.

4.1 Data Transformation

The selected dataset for this research was collected by an insurance company and contains over 500,000 trips by more than 2500 drivers. It is a time-series of GPS dataset; each record has similar characteristics. Geographic positions are denoted in sequence as Pt=(t,Xt,Yt), where t represents the time dimension, and (Xt , Yt) denotes the location of the vehicle in 2D coordinate axes. As mentioned, the structure of the trajectory data is changed so that it can be used with (Dong et al. 2016; Dong et al. 2017) deep neural network algorithms. First, the trajectory data is divided evenly into window sizes of (Ls) that shift (Ls/2). Five features are calculated in each window. These are derived from adjacent points and include speed norm, changes in speed norm, acceleration norm, changes in acceleration, and angular speed. These variables are known as the basic features. Once the basic features have been calculated, the statistical features for each segment are generated in another sliding window of length (Lf) with a shift length of (Lf/2). The statistical features include average, minimum, maximum, first, second, third quarters, and standard deviation. A generated sample of the basic features and statistical features are depicted in Figure 4. Input parameters of Ls = 256 and Lf = 4 would generate a matrix with 35 features and 128 rows in each segment as the input for the neural network model.

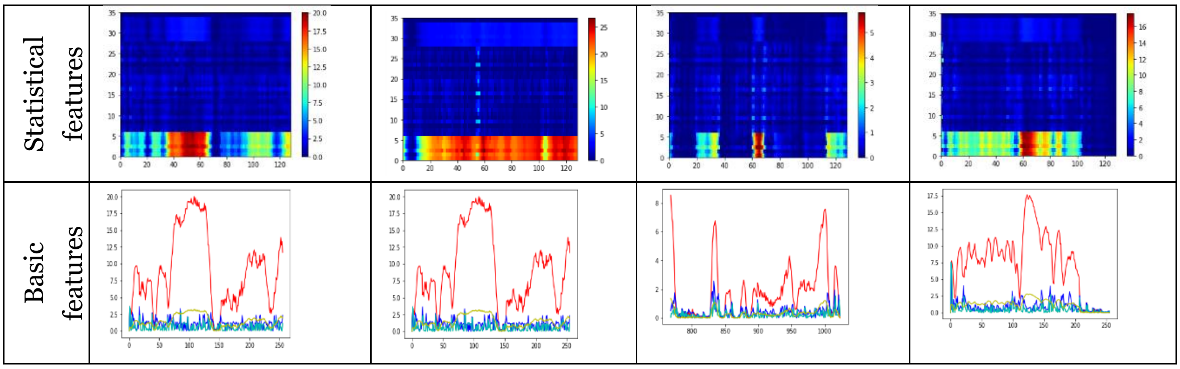

Figure 4: Transformed data from the trajectory dataset

4.2 Risk Modelling

After data transformation, the GPS data is ready for use in the deep neural network model. For this step, several deep learning algorithms need to be developed to extract driving styles from the dataset. The training sample needs to be large, and a deep neural network needs to be built from this sample to find unique driving patterns. To the best of our knowledge, the deep sparse autoencoder and recurrent neural networks would be the most suitable architectures for algorithm development. The output from these models will be the driving styles that form the criteria for our decision-making framework.

The output of the criteria mining model are the driving styles. These criteria (driving styles) are used to generate the risk model matrix proposed in Eq. (1). The matrix cells are filled by training data. Each cell represents the number of driving patterns repeated by different drivers. When the risk model's matrix is completed, the weight of the extracted styles (W_{Ci}) will be calculated with Eq. (2). These weights are numerical values for discriminating normal behaviour form dangerous.

4.3 Risk Prediction

To calculate the risk score for new drivers, unique patterns of new drivers should be detected by driving style detection component. Then, the risk score of new drivers from the testing data is calculated using Eq. (3). The weight of each criteria has been calculated in risk modelling step which is proposed in Eq. (2). This score determines the risk of drivers based on the distance of their behaviours from the normal or abnormal behaviours of other drivers.

4.4 Model Evaluation

The final step in our research is to evaluate the model. To assess the accuracy of the framework, we will compare the risk scores calculated by our model with other approaches used in this sector. These approaches also centre on risk scores, but they are a little different on the measure that define for risk calculation. For example, driver's risk score can be measured based on the mileage that they drove before, and different variations of car speed and acceleration could be another measure for comparison. Finally, the expert opinions would be the last method to judge about the performance of our models rather than other methods. We anticipate that our methodology will show superior results to these methods.

5 Conclusion

Insurers are using cutting-edge technologies to improve and upgrade existing insurance products, developing innovative new technologies, and reshaping the industry. Key transformative technologies typically bring huge datasets, but retaining that data without the tools to analyse and extract valuable insights from it is a waste of time. One such technology that has recently garnered research attention is telematics devices. This paper proposes a conceptual model for risk prediction that applies the value of telematics data while considering the characteristics and capabilities of deep learning. The model's components comprise data transformation, criteria mining, risk modelling, driving style detection, and risk prediction. The criteria mining component relies on deep learning and is responsible for extracting driving styles from the dataset. Our proposed algorithm is the deep sparse autoencoder because it has shown outstanding performance in similar problems.

In future, we will see the development of a full system based on the conceptual model that facilitates decision making for car insurance companies. Moreover, fuzzy logic could be applicable for improving our results in areas with uncertainty.

6 References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. 2014. "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on audio, speech, and language processing* (22:10), pp. 1533-1545.

Baecke, P., and Bocca, L. 2017. "The Value of Vehicle Telematics Data in Insurance Risk Selection Processes," *Decision Support Systems*).

Collobert, R., and Weston, J. 2008. "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," *Proceedings of the 25th international conference on Machine learning*: ACM, pp. 160-167.

Dong, W., Li, J., Yao, R., Li, C., Yuan, T., and Wang, L. 2016. "Characterizing Driving Styles with Deep Learning," *arXiv preprint arXiv:1607.03611*).

Dong, W., Yuan, T., Yang, K., Li, C., and Zhang, S. 2017. "Autoencoder Regularized Network for Driving Style Representation Learning," *arXiv preprint arXiv:1701.01272*).

Duri, S., Gruteser, M., Liu, X., Moskowitz, P., Perez, R., Singh, M., and Tang, J.-M. 2002. "Framework for Security and Privacy in Automotive Telematics," *Proceedings of the 2nd international workshop on Mobile commerce*: ACM, pp. 25-32.

Huang, T.-H., Nikulin, V., and Chen, L.-B. 2016. "Detection of Abnormalities in Driving Style Based on Moving Object Trajectories without Labels," *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on*: IEEE, pp. 675-680.

Husnjak, S., Peraković, D., Forenbacher, I., and Mumdziev, M. 2015. "Telematics System in Usage Based Motor Insurance," *Procedia Engineering* (100), pp. 816-825.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, pp. 1097-1105.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436-444.

Lin, N., Zong, C., Tomizuka, M., Song, P., Zhang, Z., and Li, G. 2014. "An Overview on Study of Identification of Driver Behavior Characteristics for Automotive Control," *Mathematical Problems in Engineering* (2014).

Liu, H., Taniguchi, T., Tanaka, Y., Takenaka, K., and Bando, T. 2017. "Visualization of Driving Behavior Based on Hidden Feature Extraction by Using Deep Learning," *IEEE Transactions on Intelligent Transportation Systems*).

Naderpour, M., Lu, J., and Zhang, G. 2014a. "An Intelligent Situation Awareness Support System for Safety-Critical Environments," *Decision Support Systems* (59), pp. 325-340.

Naderpour, M., Lu, J., and Zhang, G. 2014b. "A Situation Risk Awareness Approach for Process Systems Safety," *Safety Science* (64), pp. 173-189.

Niu, L., Lu, J., and Zhang, G. 2009. "Cognition-Driven Decision Support for Business Intelligence," *Models, Techniques, Systems and Applications. Studies in Computational Intelligence, Springer, Berlin*).

Paefgen, J., Staake, T., and Fleisch, E. 2014. "Multivariate Exposure Modeling of Accident Risk: Insights from Pay-as-You-Drive Insurance Data," *Transportation Research Part A: Policy and Practice* (61), pp. 27-40.

Purba, J. H., Lu, J., Zhang, G., and Pedrycz, W. 2014. "A Fuzzy Reliability Assessment of Basic Events of Fault Trees through Qualitative Data Processing," *Fuzzy Sets and Systems* (243), pp. 50-69.

Siami, M., Gholamian, M., Basiri, J., and Fathian, M. 2011. "An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy of Credit Scoring," *Model and data engineering*), pp. 133-142.

Siami, M., Gholamian, M. R., and Basiri, J. 2014. "An Application of Locally Linear Model Tree Algorithm with Combination of Feature Selection in Credit Scoring," *International Journal of Systems Science* (45:10), pp. 2213-2222.

Wahlström, J., Skog, I., and Händel, P. 2015. "Driving Behavior Analysis for Smartphone-Based Insurance Telematics," *Proceedings of the 2nd workshop on Workshop on Physical Analytics*: ACM, pp. 19-24.